Introducción a la Psicometría

Teoría Clásica de los Tests y Teoría de la Respuesta al Ítem

(Febrero 2006)

Francisco J. Abad
Jesús Garrido
Julio Olea
Vicente Ponsoda

INDICE

INTRODUCCIÓN

La Psicometría se ocupa de los problemas de medición en Psicología, utilizando la Estadística como pilar básico para la elaboración de teorías y para el desarrollo de métodos y técnicas específicas de medición. Usualmente, suelen diferenciarse varios núcleos temáticos diferentes propios de la Psicometría:

a) **Teorías de la medición.** Tienen como objetivo establecer las condiciones y propiedades de las asignaciones numéricas que pueden realizarse. El establecimiento de diferentes escalas de medida, tal como lo hizo Stevens, es un ejemplo de este primer núcleo de interés.

b) **Escalamiento.** En el terreno de la Psicofísica, e históricamente desde los trabajos de Fechner en el siglo XIX, se plantea el problema de la medición de las sensaciones que generan diferentes niveles de estimulación física. Thurstone será el responsable del tránsito del escalamiento psicofísico al escalamiento psicológico, donde se proponen modelos y técnicas para la medición de atributos estrictamente psicológicos.

c) **Teorías de los Tests.** A principios del siglo XX, Spearman propone una formulación matemática para estudiar las propiedades métricas de las puntuaciones que se asignan mediante tests, elaborados en ese tiempo (recuérdese los tests de inteligencia de Binet), para cuantificar el nivel de las personas en funciones psicológicas superiores. La principal preocupación de Spearman es incorporar en la formulación matemática los errores de medida que se cometen en la aplicación de los tests psicológicos. Este es el inicio de la Teoría Clásica de los Tests (TCT), que será descrita en 1950 en el libro de Gulliksen "Theory of Mental Tests". En la década de los 60 aparecen dos libros, uno de Rasch y otro de Lord y Novick, donde se describen los primeros desarrollos de una nueva perspectiva en el estudio de las propiedades psicométricas de los tests, la Teoría de la Respuesta al Ítem (TRI), que pretende resolver algunos de los problemas que plantea la TCT.

En otras asignaturas del plan de estudios se tratan los temas de Teorías de la medición y Escalamiento. En las siguientes páginas proporcionamos una descripción de la TCT, cuyos desarrollos siguen empleándose (en nuestro país casi de forma exclusiva) en la práctica para analizar la bondad métrica de los tests psicológicos, y una introducción a la TRI, que pensamos se irá imponiendo progresivamente, tal como ocurre en otros sitios.

La actividad profesional del psicólogo requiere en muchos momentos la utilización y/o construcción de tests que pretenden evaluar determinados constructos psicológicos no susceptibles a un proceso de medición directa. Resulta usual, por ejemplo, en el ámbito de la Psicología Educativa, la aplicación de tests de inteligencia, de hábitos de estudio, de motivación, de habilidad lectora o de intereses vocacionales. En el terreno de la psicoterapia individual, un psicólogo aplica determinadas pruebas para diagnosticar los problemas depresivos de un cliente, su estilo atribucional, la calidad de sus relaciones sexuales o su nivel de asertividad. Los psicólogos que se ocupan de la selección de personal en grandes organizaciones también utilizan tests para determinar, al menos inicialmente, cuáles son las personas del grupo de aspirantes que mejor pueden desempeñar el puesto de trabajo.

Cada vez es mayor el número de tests disponibles en el mercado para su utilización. Basta con ojear los catálogos de empresas consultoras especializadas (TEA, MEPSA, COSPA,...) para percatarnos de la gran extensión de atributos psicológicos que podemos ya medir mediante tests. El psicólogo necesita conocer las posibilidades de cada uno de estos tests: la información que aporta, cómo se interpretan las puntuaciones que proporciona, en qué grado podemos fiarnos de estas puntuaciones, para qué tipo de personas resulta apropiada su aplicación, etc. El manual de estos tests suele incluir datos empíricos sobre todos estos aspectos, que determinarán en gran parte las garantías que nos ofrece la prueba que vamos a aplicar.

Sin embargo, y debido fundamentalmente a la relativa juventud de la Psicología, los profesionales no se encuentran con todos los tests que pueden necesitar para su actividad laboral cotidiana. No resulta extraño, por ejemplo, que un psicólogo social tenga que construir un test concreto para evaluar la actitud que tiene la población de estudiantes universitarios hacia grupos marginados, que un orientador escolar necesite elaborar un test para conocer la opinión de los profesores hacia la LOGSE o que un psicólogo clínico precise de una prueba concreta para evaluar determinados aspectos de las relaciones de los adolescentes con sus padres.

Parece razonable, por tanto, y así es nuestra opinión, que un psicólogo adquiera las destrezas necesarias para valorar la información psicométrica que incluyen los tests comercializados y, además, que conozca los métodos y técnicas fundamentales para diseñar una prueba concreta con fines específicos. Trataremos de ayudarle a ello en las siguientes páginas.

En la exposición que vamos a realizar en los primeros 5 capítulos, tratamos de describir el proceso natural que se sigue en la construcción de un test, y que básicamente se resume en las siguientes fases:

1. Definición del constructo.
2. Construcción del test provisional.
3. Aplicación a una muestra.
4. Análisis de ítems.
5. Estudio de la fiabilidad del test.
6. Estudio de la validez del test.
7. Baremación.

Las cuatro primeras fases se refieren a ciertas estrategias lógicas (algunas con cierto fundamento estadístico) que nos conducen a seleccionar la forma y contenidos más apropiados del test. Las fases 5 y 6 resultan fundamentales, dado que se refieren a la comprobación empírica de las garantías psicométricas que la prueba manifiesta como instrumento de medición. Básicamente, estas garantías se refieren a su precisión (fiabilidad) y a la comprobación práctica del contenido auténtico que estamos evaluando (validez). La denominada Teoría Clásica de los Tests, cuya descripción es parte fundamental de estas páginas, permite abordar estos problemas con cierto rigor. Una vez que disponemos de la versión definitiva del test, aplicada a una muestra representativa de la población de personas a la que va dirigido, se procede a la fase de baremación, que sirve para interpretar una puntuación concreta en relación con las que obtiene la muestra seleccionada.

El último de los temas de estos apuntes pretende iniciar al alumno en los fundamentos de la **Teoría de la Respuesta al Ítem**, y será entonces cuando comentemos las diferencias principales entre ambas aproximaciones.

Estos apuntes contienen una breve descripción de los principales contenidos teóricos de la asignatura Introducción a la Psicometría. Dentro de las actividades prácticas de la asignatura, los estudiantes habrán de analizar un test de rendimiento óptimo y elaborar un test de rendimiento típico, para lo que habrán de seguir todos los pasos indicados aquí.

TEMA I: REDACCIÓN Y ANÁLISIS DE ÍTEMS

1.- INTRODUCCIÓN

Mientras que la mayoría de los atributos físicos (altura, peso, etc. ...) resultan directamente medibles, los atributos (constructos o rasgos) psicosociales resultan ser conceptualizaciones teóricas que no son accesibles a la medición directa y para los que no existen "metros" o "balanzas" diseñados para medirlos de manera precisa. La actitud hacia el aborto, el nivel de cohesión grupal, el grado de extroversión, el cociente intelectual, la postura hacia el consumo de drogas, el grado de liderazgo,...., todos ellos son constructos que deben medirse mediante instrumentos específicamente diseñados: los tests, cuestionarios o inventarios. Nadie dudaría de que un metro bien diseñado mide longitud y que lo hace de manera precisa, pero la bondad y la precisión de un cuestionario no se puede presuponer; más bien son una cuestión de grado y siempre susceptibles de mejora.

En definitiva, un cuestionario está formado por una serie de elementos o **ítems** (elementos, reactivos, preguntas, cuestiones, situaciones análogas,...) a los que cada individuo debe responder. Después de cuantificar las respuestas de una persona a los elementos del cuestionario, se pretende asignar una puntuación (a veces varias) a esa persona respecto al constructo o atributo que se pretende medir con el cuestionario, una puntuación que debería indicar el grado en que la persona participa del atributo, constructo o rasgo a evaluar.

Nos enfrentamos así a un proceso de medición indirecta que incluye la misma construcción del instrumento de medida, proceso que se inicia con la definición clara del constructo a evaluar.

2.- DEFINICIÓN DEL CONSTRUCTO

El primer paso consiste en proporcionar una definición operacional del constructo o rasgo que pretendemos medir. Por ejemplo, si hablamos de dogmatismo, debemos establecer los diversos componentes o manifestaciones del mismo: dogmatismo ante la política, ante la educación de los hijos, ante la religión, en las relaciones familiares,... Muy relacionada con esta definición operativa es la cuestión del establecimiento de los objetivos que se pretenden conseguir con el cuestionario.

También es necesario especificar el tipo de población al que va a aplicarse la prueba y las decisiones que se pretenden tomar a partir de las puntuaciones que ofrezca. Resulta muy diferente, y determinará su contenido, que un test de inteligencia se vaya a aplicar a personas de la población general o a personas con problemas intelectuales. Un cuestionario de depresión puede utilizarse con fines científicos en una investigación o para decidir el ingreso en un centro psiquiátrico de personas con problemas depresivos.

3.- CONSTRUCCIÓN PROVISIONAL DEL CUESTIONARIO

De la definición operacional del constructo y de la delimitación de sus componentes debemos llegar a establecer un conjunto de elementos o ítems (frases, preguntas, situaciones análogas, tareas, etc.) que representen estos componentes, o mejor, las conductas mediante las que se manifiestan los diversos componentes del constructo.

Si, por ejemplo, pretendemos evaluar la tolerancia hacia los grupos marginales, un ítem podría ser el siguiente:

"Deberíamos facilitar la integración de los gitanos en nuestro país"

Parece razonable suponer que una persona tolerable estaría de acuerdo con esta afirmación, mientras que otra intolerable estaría en desacuerdo.

Enrelación con la construcción de los ítems existen dos temas importantes a tener en cuenta: el **formato de respuesta** y las normas de **redacción de los ítems**.

3.1.- FORMATO DE RESPUESTA

En tests de **rendimiento óptimo** (pruebas de rendimiento y de inteligencia) se pretende medir el rendimiento máximo al que llega cada persona ante una serie de preguntas o tareas. Usualmente, el formato de respuesta de estos ítems se ajusta a uno de los siguientes tres formatos:

a) **Elección binaria:** De dos alternativas, se elige la que se considera correcta (Sí o No; verdadero-falso).

Por ejemplo, un ítem de un test de rendimiento en Historia Moderna puede ser:

"Pi y Margall fue uno de los presidentes de la 1ª República Española" V F

b) **Elección múltiple:** Entre más de dos alternativas se elige la que se considera correcta. Es sin duda el formato de respuesta más utilizado, entre otras por razones de objetividad y otras de tipo operativo.

Por ejemplo, un ítem de un test de aptitud verbal puede ser:

"Automóvil es a volante como bicicleta es a ...."

a) Pedal
b) Sillín
c) Manillar
d) Parrilla

c) **Emparejamiento:** Consiste en encontrar las parejas entre dos conjuntos de conceptos. Por ejemplo, un ítem de un cuestionario sobre conocimientos de políticos españoles contemporáneos puede ser:

"Enlace mediante una línea el nombre del político con el partido político al que pertenece"

J. A. Durán i Lleida	*PSOE*
Carlos Solchaga	*CIU*
Iñaki Anasagasti	PNV
Rodrigo Rato	*PP*

Mediante las pruebas de **rendimiento típico** se quiere reflejar el comportamiento ordinario de las personas, no teniendo sentido el concepto de rendimiento máximo dado que el objeto de la evaluación es algún tipo de opinión, actitud o rasgo de personalidad. El formato de respuesta de los cuestionarios de rendimiento típico se ajusta a alguno de los siguientes:

a) **Opción binaria:** La persona debe manifestar si está de acuerdo o en desacuerdo con una afirmación. Por ejemplo, un ítem de un cuestionario sobre la actitud de los padres hacia los profesores de sus hijos puede ser:

"En realidad, los profesores en el colegio hacen poco más que cuidar a nuestros hijos cuando nosotros trabajamos"

Acuerdo ( ) Desacuerdo ( )

b) **Categorías ordenadas:** El formato establece un continuo ordinal de más de dos categorías, que permite a la persona matizar mejor su respuesta. Normalmente, este continuo está formado por 5 ó 7 categorías ordenadas, con una categoría central para indicar la valencia neutra y a partir de la cual posicionarse en uno u otro sentido. Por ejemplo, un ítem sobre la actitud de los adolescentes hacia el consumo de drogas, podría ser el que sigue:

"Las drogas pueden realmente resolver problemas de uno mismo"

( ) Muy en Desacuerdo
( ) Bastante en Desacuerdo
( ) Neutral
( ) Bastante de Acuerdo
( ) Muy de Acuerdo

A veces, se establecen nominalmente los dos extremos del continuo, dejando señaladas las restantes categorías del mismo:

MD ____ ____ ____ ____ ____ MA

o se ordenan numéricamente las categorías sucesivas:

1 2 3 4 5 6 7

c) **Adjetivos bipolares:** Este formato es típico de lo que se denomina "diferencial semántico", un instrumento formado por pares de adjetivos opuestos, cada uno de los cuales representa un continuo bipolar con varias categorías, y que permite estudiar el significado semántico que se atribuye a determinados constructos, personas o instituciones.

Por ejemplo:

Alegre	_____	_____	_____	*Triste*
Listo	_____	_____	_____	*Tonto*
Simpático	_____	_____	_____	*Antipático*
Feliz	_____	_____	_____	*Infeliz*
Social	_____	_____	_____	*Asocial*

3.2.- REDACCIÓN DE ÍTEMS

Algunas de las recomendaciones generales en la redacción de ítems en pruebas de **rendimiento óptimo** son las siguientes:

- La idea principal del ítem debe estar en el enunciado.
- Simplicidad en el enunciado.
- Evitar los conocimientos excesivamente triviales o excesivamente "rebuscados".
- Evitar dar información irrelevante en el enunciado.
- Evitar dar indicios sobre la solución.
- Evitar cuestiones sobre opiniones.
- No encadenar unos ítems con otros.
- Anticipar la dificultad e incluir preguntas de todo rango de dificultad (casi siempre conviene más preguntas de dificultad media).
- La dificultad no debe estar en la comprensión del ítem.
- Minimizar el tiempo de lectura.
- Evitar el uso de negaciones (si se incluyen, subrayarlas), errores gramaticales y ortográficos.

En cuanto al número de opciones, con dos distractores es suficiente; pero si la prueba es corta, es necesario un mayor número de distractores para evitar los efectos de los aciertos aleatorios. Todos los distractores deben ser de longitud y lenguaje parecidos y también se deben evitar los solapamientos entre ellos. Por supuesto, se deben evitar los llamados "ítems defectuosos" que son aquellos ítems con más de una respuesta correcta; aunque parezca absurdo son errores que se siguen cometiendo con excesiva frecuencia. Por otro lado, se deben evitar las opciones del tipo "no lo sé", "todas las anteriores son correctas" o "ninguna de las anteriores es correcta"; así como balancear la posición de la opción correcta en las diferentes preguntas para que no se sitúe siempre en la misma opción.

En las pruebas de rendimiento óptimo es muy importante tener en cuenta la dificultad existente en crear las alternativas incorrectas, dado que no deben ser posibilidades absurdas de respuesta que se puedan eliminar con cierto grado de sentido común. Bien al contrario, esas alternativas no ciertas deben ser elegidas entre los errores o confusiones que usualmente tienen las personas que no conocen la respuesta correcta de la pregunta en cuestión. Deben estar escritas en lenguaje técnico y ser plausibles para quien no conoce la respuesta, evitando en todo momento alternativas "graciosas" u otras que no serían elegidas por nadie. Otra buena recomendación en este sentido sería el uso de alternativas de respuesta que son verdaderas para otras preguntas incluidas en el cuestionario. Desde luego, el establecimiento de alternativas múltiples exige un claro conocimiento tanto del contenido a evaluar como de las personas a las que va dirigida la prueba. Una reciente revisión de las orientaciones a seguir a la hora de escribir ítems de opción múltiple se encuentra en Haladyna, Downing y Rodríguez (2002).[1]

Otras recomendaciones a tener presente en las pruebas de rendimiento óptimo son:

- El número de preguntas debe ser proporcional a la importancia dada a cada tema.
- Corregir los aciertos obtenidos por azar.
- Cuantos más ítems, mejor.

Respecto a la manera de formular las cuestiones en tests de **rendimiento típico** (declaraciones o afirmaciones ante las cuales se debe opinar), se han propuesto algunas sugerencias que pueden ayudar a su correcta redacción:

- Utilizar el tiempo presente.
- Deben ser "relevantes", en el sentido de que su contenido debe relacionarse claramente con el rasgo.
- Contenido claro, evitando excesiva generalidad. Frases cortas, simples e inteligibles. Evitar incluir dos contenidos en un ítem.
- Tener en cuenta que lo que se dice en la declaración pueda ser asumido por alguien, y no por todos.
- En escalas de actitudes, no plantear la existencia o no de hechos, sino el posicionamiento personal sobre la afirmación. Redactar ítems que discriminen entre los de actitud positiva y los de actitud negativa.
- Para minimizar la aquiescencia (tendencia a responder afirmativamente, independientemente del contenido por el que se pregunta) conviene redactar ítems de modo directo e inverso (declaraciones tanto en sentido positivo como en sentido negativo). A posteriori, se puede comprobar cómo una persona ha respondido a los ítems directos e inversos. También puede comprobarse que la correlación entre ambos tipos de ítems es alta.
- Evitar el uso dobles negaciones (no, ninguno, nunca,...) y de universales (todo, siempre, nada,...).
- En lo posible, aunque no es fácil, se debe minimizar la posibilidad de deseabilidad social (emitir respuestas socialmente aceptables para transmitir una imagen positiva). Puede deberse a varias cosas: desajuste psicológico, insinceridad consciente,... El grado de deseabilidad social que manifiestan los ítems puede evaluarse mediante

[1] Haladyna, T.M., Downing, S.M. & Rodríguez, M.C. (2002). A review of multiple–choice item writing guidelines for classroom assessment. *Applied Measurement in Education* 15, 309-334.

jueces, y comparar las respuestas de una persona con estas valoraciones. En tests de personalidad puede incluirse una escala de sinceridad.

El número de categorías que se suelen incluir en este tipo de ítems es usualmente de cinco ya que, a partir de ese número de categorías no mejoran las propiedades psicométricas de los ítems. Además, un número muy elevado de categorías (siete u ocho) lleva a inconsistencias en las respuestas, que es una fuente de error. Un número muy reducido (dos ó tres) lleva a poca discriminación (menor variabilidad) y a reducir la fiabilidad, aunque siempre puede compensarse con un mayor número de ítems. No obstante, en poblaciones especiales (niños, discapacitados, mayores…) se aconseja el uso de un menor número de categorías.

También se ha planteado en ítems de rendimiento típico (principalmente en escalas de actitudes o tests de personalidad donde no se pide el grado de frecuencia de un comportamiento) si es correcto o no la inclusión de una categoría central en las opciones de respuesta ("indiferente", "neutral", "dudo", "no sé"…). Podrían generar problemas ya que muchas veces son elegidas por aquellas personas que no se comprometen con lo que se les está preguntando, que el enunciado les resulta ambigüo o simplemente que ignoran el contenido del enunciado. En realidad, deberían ser seleccionadas por las personas auténticamente indecisas. La investigación en este sentido nos dice que los indicadores psicométricos de los ítems no se alteran mucho con o sin categoría central, cuando el número de categorías es mayor de tres. En todo caso, se puede comprobar si las personas con nivel medio en el total del test tienden a elegir más frecuentemente las categorías centrales.

4.- CUANTIFICACIÓN DE LAS RESPUESTAS

Una vez establecido el formato de respuesta que se considera más apropiado para el caso, y de cara al estudio psicométrico de la prueba, es preciso decidir la manera de cuantificar los posibles resultados a las cuestiones. En general, para los ítems de cuestionarios de rendimiento óptimo se cuantificará con 1 el acierto y con 0 el error, de tal manera que la puntuación directa de un sujeto en un cuestionario determinado será igual al número de ítems que ese sujeto acierta.

La cuantificación de las respuestas a ítems de pruebas de rendimiento típico requiere ciertos matices. Dado un formato de respuesta determinado (opción binaria, categorías ordenadas o adjetivos bipolares) es necesario cuantificar las posibles respuestas a un ítem teniendo en cuenta que la alternativa con mayor valor sea la que indique mayor nivel de rasgo, aptitud y opinión.

Por ejemplo, para un ítem con formato de respuesta de opción binaria (acuerdo/desacuerdo) puede cuantificarse el acuerdo como 1 y el desacuerdo como 2, o viceversa. Depende de que el ítem esté planteado para medir de manera directa o inversa el constructo de interés. Estos pueden ser 2 ítems de un cuestionario de actitud ante al aborto voluntario:

Ítem A: "Abortar es matar".

De acuerdo ( ) En desacuerdo( )

Ítem B: "El bienestar de la madre también importa".
De acuerdo ( ) En desacuerdo ( )

En el ítem A, el acuerdo se puntuaría con 1 y el desacuerdo con 2, ya que estar en desacuerdo con esa afirmación indica una actitud más positiva hacia el aborto voluntario. En el ítem B, sin embargo, el acuerdo se puntuaría con 2 y el desacuerdo con 1, ya que estar de acuerdo con esa afirmación indica una actitud más positiva hacia el aborto.

Si el formato de respuesta es de "n" categorías ordenadas, las diversas categorías se cuantificarán normalmente desde 1 hasta n, teniendo en consideración (como en el caso anterior) la dirección de la afirmación o cuestión. Por ejemplo, para 5 categorías, las dos posibles cuantificaciones serán:

Muy en desacuerdo	*Bastante en desacuerdo*	*Neutral*	*Bastante de acuerdo*	*Muy de acuerdo*
1	*2*	*3*	*4*	*5*
5	*4*	*3*	*2*	*1*

También se puede asignar el 0 a la categoría central, valores negativos a las categorías que se encuentran a la izquierda y positivos a las que se encuentran ubicada a la derecha.

En estos casos, la puntuación directa de un sujeto en un test (o subtest) resulta de sumar las cantidades asignadas por el constructor de la prueba a las diferentes respuestas que el sujeto ha emitido; según esto, convendría cuantificar las diversas alternativas con valores entre 1 y n para evitar una puntuación directa negativa.

5.- ANÁLISIS DE ÍTEMS

Los ítems o cuestiones se han formulado de manera lógica para que midan (y lo hagan bien) el constructo, variable, o rasgo que interesa evaluar con el cuestionario. Ahora bien, el grado en que cada ítem es un "buen medidor" del rasgo de interés es algo que se puede comprobar estadísticamente de manera sencilla si obtenemos tres indicadores para cada ítem:

a) El índice de dificultad.

b) El índice de homogeneidad.

c) El índice de validez.

Para ello, tras aplicar el cuestionario provisional a una muestra de sujetos representativa de la población a la que va dirigida la prueba (se aconseja entre 5 y 10 veces más sujetos que ítems), y una vez cuantificadas las respuestas de cada individuo, se forma una matriz de datos de sujetos x ítems:

	Ítems			
	1	2	3..... n	X
Sujeto nº 1				
Sujeto nº 2				
Sujeto nº 3				
.				
.				
.				
.				
.				
.				
Sujeto nº N				

Un elemento a_{ij} de esta matriz indica el valor asignado a la respuesta que da el sujeto i al ítem j. Sumando por filas podemos obtener las puntuaciones directas (X) de los sujetos en el total del test.

Veamos cómo se obtienen (y qué sentido tiene su obtención) los tres índices citados anteriormente.

5.1.- ÍNDICE DE DIFICULTAD

Este primer indicador sirve para cuantificar el grado de dificultad de cada cuestión, por lo que sólo tiene sentido su cálculo para ítems de tests de rendimiento óptimo.

El índice de dificultad de un ítem j se define como el cociente entre el nº de sujetos que lo han acertado (A_j) y el nº total de sujetos que lo han intentado resolver (N_j)

$$D_j = \frac{A_j}{N_j}$$

Atendiendo a la disposición de datos en la matriz expuesta más arriba, el índice de dificultad de un ítem (columna) j será el cociente entre el nº de unos y el total de unos y ceros que tiene la columna. Los sujetos que han omitido el ítem (no han contestado) no se contabilizan en N_j.

Ejemplo: Supongamos que la siguiente tabla recoge las respuestas de una muestra de 10 personas a un test formado por 6 ítems dicotómicos (1 indica acierto y 0 error):

		Ítems						
		1	2	3	4	5	6	X
Sujetos	1	0	0	0	1	1	1	3
	2	0	1	-	0	-	1	2
	3	0	0	1	-	0	1	2
	4	0	0	0	-	1	1	2
	5	0	1	0	1	-	1	3
	6	0	1	-	-	-	1	2
	7	0	0	-	1	1	1	3
	8	0	0	1	-	0	-	1
	9	0	1	0	-	0	1	2
	10	0	1	0	-	0	1	2
	A_j	0	5	2	3	3	9	
	N_j	10	10	7	4	7	9	
	D_j	0	0.5	0.29	0.75	0.43	1	

Con estos resultados podemos comprobar varios aspectos de la interpretación de D_j:

- El valor mínimo que puede asumir D_j es 0 (ningún sujeto acierta el ítem) y el valor máximo 1 (todos los sujetos que lo intentan lo aciertan).

- A medida que D_j se acerca a 0 indica que el ítem ha resultado muy difícil; si se acerca a 1, que ha resultado muy fácil; y si se acerca a 0,5, que no ha resultado ni fácil ni difícil.

- D_j está relacionado con la varianza de los ítems: Si D_j es 0 ó 1, la varianza es igual a cero; a medida que D_j se acerca a 0,5, la varianza del ítem aumenta. De nada sirve un ítem con $D_j = 0$ o $D_j = 1$, ya que no discriminaría entre los diferentes sujetos (todos aciertan o todos fallan).

Al diseñar un cuestionario de rendimiento óptimo, al inicio se sitúan los ítems más fáciles (con mayor D_j); en la parte central, los de dificultad media (entre 0,30 y 0,70); y al final, los más difíciles (con menor D_j). El número de ítems de cada categoría de dificultad que deben incluirse en el test depende de los objetivos que quiera conseguir la persona que diseña el cuestionario. En general, la mayor parte de los ítems deben ser de dificultad media.

5.2.- ÍNDICE DE HOMOGENEIDAD

El índice de homogeneidad, llamado a veces índice de discriminación, de un ítem (H_j) se define como la correlación de Pearson entre las puntuaciones de los N sujetos en el ítem j y las puntuaciones X en el total del test:

$$H_j = r_{jx}$$

Según la disposición de la matriz de datos, para obtener los H_j de los ítems, debemos calcular la correlación entre las columnas j y la columna X de puntuaciones directas en la prueba.

Ejemplo: Supongamos un test formado por 3 ítems con formato de respuesta de categorías ordenadas, que se valoran entre 0 y 5. Después de aplicarse a un grupo de 5 sujetos se obtienen los siguientes datos:

Ítems

		1	2	3	X
Sujetos	1	2	3	5	10
	2	3	1	0	4
	3	5	4	5	14
	4	0	1	0	1
	5	4	3	0	7

Puede comprobarse que los índices de homogeneidad de los 3 elementos son:

$H_1 = r_{1x} = 0{,}75$
$H_2 = r_{2x} = 0{,}94$
$H_3 = r_{3x} = 0{,}86$

El índice de homogeneidad de un ítem nos va a informar del grado en que dicho ítem está midiendo lo mismo que la prueba globalmente; es decir, del grado en que contribuye a la homogeneidad o consistencia interna del test. Los ítems con bajos índices de homogeneidad miden algo diferente a lo que refleja la prueba en su conjunto. Si con el test se pretende evaluar un rasgo o constructo unitario, deberían eliminarse los que tienen un H_j próximo a cero.

En ocasiones, un test está formado por diferentes subtests con contenidos distintos. En este caso, los H_j deben obtenerse con relación a las puntuaciones directas del subtest concreto.

Cuando un H_j es negativo y alto, debemos cuestionar el sistema de cuantificación de las respuestas que se ha seguido en ese ítem. Si un ítem obtiene una correlación negativa y alta con el total de la prueba, seguramente es debido a que se ha cuantificado erróneamente el ítem (se ha tomado como directo siendo inverso, o viceversa).

Cuando un test tiene un número pequeño de ítems, resulta más apropiado obtener el **índice de homogeneidad corregido** ($r_{j.x-j}$). Consiste en correlacionar las puntuaciones en un ítem con las puntuaciones en el total del test después de restar de este total las puntuaciones del ítem cuyo índice queremos obtener. En el ejemplo precedente, el índice de homogeneidad corregido para el ítem 1 será 0.49, resultado de correlacionar la 1ª columna de la tabla (2, 3, 5, 0, 4) con la columna (10-2 = 8, 4-3 = 1, 14-5 = 9, 1-0 = 1, 7-4 = 3). Análogamente, los índices de homogeneidad corregidos para los ítems 2 y 3 son, respectivamente, 0.89 y 0.54. Como resulta lógico suponer, el H_j corregido de un ítem suele ser inferior a su H_j sin corregir.

5.3.- ÍNDICE DE VALIDEZ

Las puntuaciones de los N sujetos en un ítem j pueden correlacionarse también con las que estos sujetos obtienen en un criterio de validación externo al test (Y); esta correlación define el índice de validez del ítem j:

$$V_j = r_{jy}$$

El criterio de validación "Y" es una medida diferente del test para reflejar el mismo rasgo u otro muy relacionado, de tal manera que si el test mide lo que se pretende, debería correlacionar de forma elevada con el criterio. Por ejemplo, un criterio para validar un test de inteligencia verbal puede ser otro test que incluye cuestiones verbales; los supervisores de unos trabajadores podrían valorar el grado de motivación de cada uno y utilizar estas valoraciones como el criterio de validación de un test de motivación laboral; el total de ventas en pesetas que realizan los vendedores puede ser un buen criterio para validar un test de aptitud para la venta.

Supongamos que partimos de los datos del ejemplo precedente, y que conocemos las puntuaciones directas de las 5 personas en un criterio Y:

Sujeto: 1 2 3 4 5

Y : 5 3 6 0 6

Los índices de validez de los tres ítems serán:

$$V_1 = r_{1Y} = 0{,}87$$

$$V_2 = r_{2Y} = 0{,}88$$

$$V_3 = r_{3Y} = 0{,}54$$

Los elementos que tengan una correlación con el criterio próxima a cero deberían eliminarse de la prueba, en la medida que no contribuyen a evaluar el rasgo que se pretende medir. Si lo

que se pretende es seleccionar los ítems que más contribuyen a la validez del cuestionario, de entre los ítems de igual varianza, serían preferibles los que tienen alto V_j y bajo H_j.

6.- ANÁLISIS DE OPCIONES INCORRECTAS DE RESPUESTA

Muy en relación con el análisis de ítems se encuentra el tema del estudio de los patrones de respuesta que se han dado a las diferentes alternativas de cada ítem. Para un ítem concreto de una prueba de rendimiento óptimo, lo ideal es que la alternativa seleccionada en mayor medida sea la correcta; cada una de las alternativas incorrectas del ítem debe también ser seleccionada por un número de personas que, aun siendo inferior al que selecciona la alternativa correcta, ratifique como adecuadas (como bien planteadas) dichas alternativas incorrectas.

Observemos los siguientes porcentajes de respuesta obtenidos en las diferentes opciones de tres ítems de un determinado test:

Ítem	Opción correcta	Porcentaje de respuesta				
		a	b	c	d	e
1	b	16	40	15	14	15
2	c	35	15	21	17	12
3	a	60	1	21	18	0

El patrón de respuestas obtenido para el ítem 1 es adecuado, pues la mayor parte de la muestra selecciona la alternativa correcta, mientras que las incorrectas son seleccionadas por un porcentaje parecido de personas. El ítem 2 seguramente no es muy adecuado, pues la muestra selecciona en mayor grado una alternativa incorrecta como la buena; al menos, habría que reformular esa alternativa incorrecta. Para el ítem 3, los problemas se refieren a dos alternativas incorrectas que apenas si son seleccionadas por la muestra; también habría que reformular esas dos opciones de respuesta.

7.- CORRECCIÓN DE LOS EFECTOS DEL AZAR

En los tests formados por ítems de opción múltiples de las que sólo una es correcta, podemos sobrestimar la puntuación directa de una persona dado que alguno de sus aciertos ha podido producirse por azar. El problema entonces consiste en establecer un procedimiento para descontar del número total de aciertos (A) los que se han producido por azar (A_a).

Si asumimos que, cuando no se conoce la respuesta correcta a un ítem, todas las alternativas de respuesta son equiprobables, la probabilidad de acertar al azar ese ítem se puede establecer como:

$$P(A_a) = 1/n$$

siendo *n* el número de alternativas del ítem.

De la misma forma, la probabilidad de errar el ítem será:

$$P(E) = 1 - (1/n) = (n-1) / n$$

Llamemos R_a el nº de respuestas aleatorias que proporciona (es decir, el número de ítems que ha contestado sin saber la solución). De las R_a, algunas serán aciertos aleatorios (A_a) y otras serán errores (E). Nuestro objetivo es estimar los A_a para descontarlos del número total de aciertos que ha tenido en realidad la persona. Lo haremos de la siguiente forma:

El nº total de errores se puede establecer como el producto del valor R_a por la probabilidad de cometer un error:

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

Si despejamos R_a de esta expresión, podremos estimarla a partir de datos conocidos (E y n):

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

Siguiendo el mismo razonamiento, el número de aciertos aleatorios se puede estimar multiplicando el valor R_a por la probabilidad de cometer un acierto por azar (A_a):

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

Si realizamos las sustituciones oportunas, podemos llegar a estimar A_a:

$$A_a = \frac{n}{n-1} E \frac{1}{n} = \frac{1}{n-1} E$$

Esta va a ser la fórmula para estimar A_a, a partir de los errores cometidos y del número de alternativas que tienen los ítems. Podemos observar que cada error se pondera por la expresión 1/(n-1), lo que significa que por cada error hay que descontar el resultado de ese

cociente: en tests de 2 alternativas de respuesta, hay que descontar 1 punto por cada error; en tests de 3 alternativas, hay que descontar 0,5 por cada error; en tests de 4 alternativas, hay que descontar 0,33 puntos por cada error; y así sucesivamente.

La puntuación directa corregida de una persona en el test se obtiene entonces haciendo:

$$X_c = A - A_a$$

Ejemplo: Un test de conocimientos del idioma inglés está formado por 140 ítems con 5 opciones de respuesta cada uno. A continuación se detallan el nº de aciertos (A), errores (E) y omisiones (O) que obtuvieron 3 personas:

Persona	A	E	0
1	112	28	0
2	110	12	18
3	109	0	31

Si atendemos únicamente al número de aciertos obtenidos, parece claro que quien más inglés sabe es la persona 1, seguida de la 2 y en último lugar la persona 3. Sin embargo, corrigiendo los efectos del azar, obtenemos las puntuaciones directas corregidas siguientes:

$$Xc_1 = 112 - \frac{28}{4} = 105$$

$$Xc_2 = 110 - \frac{12}{4} = 107$$

$$Xc_3 = 109 - \frac{0}{4} = 109$$

Podemos comprobar que la corrección afecta sensiblemente al orden que establecemos respecto al dominio del idioma inglés. Además, si nos fijamos en la corrección hecha para la persona 3, vemos que no se le ha descontado nada; esto es debido a que no cometió ningún error.

EJERCICIOS

1. A continuación se expone una escala de **actitud favorable ante las drogas**. Cada frase se responde con "N" (nunca), "PV" (pocas veces), "AV" (a veces), "MV" (muchas veces) o "S" (siempre).

a) A menudo me influyen más las opiniones de los demás que las mías propias . . . _____
b) Evito vivir situaciones límites . . . . . _____
c) No me importaría tomar estimulantes para disminuir la sensación de fatiga física o mental en el trabajo . . . . . _____
d) Me considero capaz de resolver un problema por mi mismo . . . . . _____
e) Me gustaría decir "NO", pero no puedo . . . . . _____

Las respuestas de 4 personas a la escala han sido las siguientes:

	ítem a	ítem b	ítem c	ítem d	ítem e
sujeto 1	S	PV	MV	N	S
sujeto 2	PV	MV	AV	MV	AV
sujeto 3	N	S	N	MV	N
sujeto 4	MV	N	AV	MV	PV

A partir de la información anterior, complete la siguiente tabla de datos. Para ello deberá obtener las puntuaciones en cada ítem y en el total de la escala:

	ítem a	ítem b	ítem c	ítem d	ítem e	TOTAL
sujeto 1						
sujeto 2						
sujeto 3						
sujeto 4						

2. Diga si las siguientes afirmaciones referidas al índice de dificultad (D_j) son verdaderas o falsas.

a) Sólo tiene sentido su cálculo en pruebas de rendimiento óptimo.
b) Se deben seleccionar sólo aquellos ítems con D_j próximos a 1.
c) Si un ítem tiene una alta varianza, su índice de dificultad será alto.
d) A un ítem de baja varianza le corresponde necesariamente un índice de dificultad bajo.

3. A continuación se ofrece una matriz ítems por sujetos:

a) ¿Cuál es el ítem más fácil?
b) ¿Cuál es el más difícil?
c) ¿Cuál es el ítem en el que las personas muestran más variabilidad?
d) ¿Cuál es en el que muestran menos variabilidad?

	ítem 1	ítem 2	ítem 3	item 4	ítem 5	ítem 6
sujeto 1	1	1	0	1		
sujeto 2	1	1	1	1	1	0
sujeto 3	0	1	0	0	0	0
sujeto 4	1	1	0	0		
sujeto 5	1	1	0	1	0	0
sujeto 6	0	1	1	0	0	0

4. Responda a las siguientes afirmaciones indicando si lo que se dice es verdadero o falso. Justifique sus respuestas.

a) El índice de homogeneidad de un ítem indica en que grado mide lo mismo que el test.
b) Un ítem con un H_j bajo siempre debe ser descartado en un proceso de selección.
c) El índice de homogeneidad permite ver en qué medida un ítem permite predecir un criterio.
d) Cuando construimos un cuestionario que mide varios rasgos debemos rechazar aquellos ítems que correlacionen poco con la puntuación total en el test.
e) Un ítem con un índice de homogeneidad alto pero con un bajo índice de validez no es necesariamente un mal ítem. Estos resultados pueden deberse a que el criterio seleccionado sea poco adecuado.

5. Un test tiene 3 ítems dicotómicos y su media es 1.7. Sabemos que no se han dejado ítems sin responder y que

	ítem 1	ítem 2	ítem 3
D_j	?	?	0.8
S_j^2	0.25	?	?
H_j	0.6	0.4	0.2
V_j	0.4	0.3	0.5

a) Complete la tabla.
b) Atendiendo al índice de dificultad, ¿cuál es el peor ítem?
c) Atendiendo al índice de homogeneidad, ¿cuál es el peor ítem?
d) Atendiendo exclusivamente al índice de validez, ¿cuál es el peor ítem?

6. Se ha construido una pequeña prueba de 6 elementos de Verdadero-Falso, para realizar una primera valoración de la rapidez visomotora de las personas que desean obtener el carnet de conducir. Un grupo de 10 personas respondió al test y a una prueba de agilidad psicomotora, que se consideró como un criterio adecuado de validación. La siguiente tabla recoge las respuestas del grupo a los elementos del test y sus puntuaciones en el criterio.

	ítem 1	ítem 2	ítem 3	ítem 4	ítem 5	ítem 6	Y
sujeto 1	1	1	0	1	1	1	12
sujeto 2	1	1	1	0	1	1	11
sujeto 3	1	0	0	1	0	1	7
sujeto 4	1	0	1	1	1	0	8
sujeto 5	0	1	0	0	0	0	4
sujeto 6	1	1	0	0	1	1	10
sujeto 7	1	0	1	1	0	0	7
sujeto 8	0	0	1	1	1	1	10
sujeto 9	1	1	0	1	1	1	11
sujeto 10	1	1	1	1	1	1	12

a) Diga cuál es el ítem con mayor índice de dificultad.
b) Obtenga un indicador del grado en el que el elemento 2 mide lo mismo que la prueba.
c) Obtenga la puntuación directa corregida para la persona 8.
d) Obtenga el grado en que el ítem 5 mide lo mismo que el criterio Y.

7. Los indicadores de cuatro ítems dicotómicos han sido los siguientes:

	ítem 1	ítem 2	ítem 3	ítem 4
D_j	0,4	0,8	0,3	0,6
H_j	0,1	0,5	0,8	0,4
V_j	0,2	0,1	0,6	0,3

a) El ítem que menos contribuye a que el test de 4 ítems mida un solo rasgo es el número ____ porque _____.

b) El ítem que menos contribuye a la validez del test de 4 ítems es el número ____ porque _____.

c) El ítem que menos contribuye a la varianza del test de cuatro ítems es el número _____ porque _____.

8. A continuación se ofrecen ciertos datos de un ítem dicotómico: su índice de dificultad, varianza, índice de homogeneidad e índice de homogeneidad corregido. Identifíquelos.

0.15 es _____
0.24 es _____
0.40 es _____
0.53 es _____

9. En un test de rendimiento óptimo, un ítem tiene 4 posibles respuestas y ha sido respondido por 350 personas. 100 personas han elegido cada una de las alternativas incorrectas y 50 personas, la correcta.
a) ¿Es un ítem adecuado o debería ser modificado?
b) ¿Cuánto vale su índice de dificultad?
c) ¿Cuanto vale su varianza?

10. Una persona completa un test de 50 ítems. Acierta 30 y falla 4. Su puntuación corregida (para eliminar posibles aciertos por azar) ha sido 29 puntos. ¿Cuántas alternativas tiene cada ítem?

11. Un examen consta de 25 preguntas verdadero-falso, que se han puntuado como "0" o "1". A continuación se ofrecen las puntuaciones sin corregir (X) y corregidas para eliminar los posibles aciertos por azar (X_c) de cinco personas en el examen. Diga razonadamente qué personas han dejado preguntas sin contestar.

	X	X_c
sujeto 1	20	18
sujeto 2	15	5
sujeto 3	25	25
sujeto 4	17	12
sujeto 5	23	22

12. Creamos un test para medir conocimientos sobre el código de la circulación. Los ítems son de opción múltiple con 3 opciones de las que sólo una es correcta. Las medias de tres ítems del test han sido las siguientes: 0.1 (ítem 1), 0.6 (ítem 2) y 1 (ítem 3). Responda razonadamente.
a) ¿Qué ítem es más difícil?
b) ¿Es posible que la mitad de la muestra haya fallado simultáneamente los dos primeros ítems?

c) Sabiendo que en ninguno de los ítems ha habido omisiones, ¿Cuánto vale la varianza del ítem de más varianza de los tres?
d) A Laura le ha correspondido en el test una puntuación sin corregir de 20 y una puntuación tras corregir los aciertos por azar de 16 ¿Cuántos errores ha cometido?

13. Una muestra de 200 personas responde a un test de rendimiento óptimo de tres alternativas. La siguiente tabla muestra las personas que eligieron cada alternativa en cada ítem, y cual es en cada uno la alternativa correcta.

	Alternativa "a"	Alternativa "b"	Alternativa "c"	Alternativa correcta
Ítem 1	30	80	90	a
Ítem 2	140	0	60	a
Ítem 3	90	10	100	c
Ítem 4	70	80	50	b
Ítem 5	60	50	90	c

a) Sabiendo que no hubo omisiones en ninguno de los ítems, calcule la media del ítem 1.
b) ¿Cuál es el ítem más difícil? Razone su respuesta.
c) A partir del estudio de las alternativas incorrectas ¿algún ítem debería ser modificado? Razone su respuesta

14. Un test de 12 ítems está formado por 2 escalas que miden constructos distintos. La escala 1 está integrada por los primeros 4 ítems y la escala 2 por los últimos 8 ítems. Las siguientes dos tablas muestran los índices de homogeneidad (H) y homogeneidad corregidos (HC) de los tres primeros ítems en relación al test de 12 ítems y en relación a la escala 1.

Tabla 1	Ítem 1	Ítem 2	Ítem 3
H	0.572	0.454	0.575
HC	0.456	0.281	0.437

Tabla 2	Ítem 1	Ítem 2	Ítem 3
H	0.562	0.622	0.611
HC	0.237	0.205	0.233

Diga **razonadamente** qué tabla contiene los H y HC de los tres ítems en relación al test completo.

SOLUCIONES

1.

	ítem a	item b	item c	ítem d	ítem e	TOTAL
sujeto 1	5	4	4	5	5	23
sujeto 2	2	2	3	2	3	12
sujeto 3	1	1	1	2	1	6
sujeto 4	4	5	3	2	2	16

2. a) Verdadero
b) Falso
c) Falso
d) Falso

3. Tal y como se desprende de la tabla siguiente:
a) El ítem más fácil es el número 2, ya que todos los sujetos lo aciertan.
b) El ítem más difícil es el número 6, ya que nadie lo acierta.
c) El ítem en el que hay más variabilidad es el número 4, porque presenta la mayor varianza.
d) Los ítems de menos variabilidad son los números 2 y 6, porque la varianza es nula en ambos.

	ítem 1	ítem 2	ítem 3	ítem 4	ítem 5	ítem 6
D_j	0,67	1	0,33	0,5	0,25	0
S_j^2	0,22	0	0,22	0,25	0,19	0

4. a) Verdadero, dado que es una correlación entre las puntuaciones en el ítem y en el test.
b) Falso. Siempre que se pretenda medir un único rasgo con el test, debe ser descartado; si se pretenden medir varios rasgos, podría ser admitido.
c) Falso, la afirmación hace referencia al índice de validez.
d) Falso. Al diseñar un test que mida varios rasgos, se pretende buscar ítems que correlacionen con los ítems que miden el mismo rasgo, y que además no correlacionen con otros ítems que miden un rasgo diferente. En esta situación, la correlación entre los ítems y las puntuaciones del test pueden ser bajas.
e) Verdadero. El ítem mide lo mismo que el test, pero no mide lo mismo que el criterio, que podría ser poco adecuado.

5. a) $D_1 = 0,5$ $D_2 = 0,4$ $S_2^2 = 0,24$ $S_3^2 = 0,16$
b) Los 3 son buenos, pero el que menos varianza tiene es el 3 y, en ese sentido, es algo peor.
c) El ítem 3.
d) El ítem 2.

6. a) El ítem 1: $D_1 = 0,8$
b) $H_2 = 0.305$
c) La persona número 8: $X_c = 2$
d) $V_5 = 0.84$

7. a) El ítem 1 (menor H).
b) El ítem 2 (menor V-H).
c) El ítem 2 (D más distante de 0.5).

8. Por ser un ítem dicotómico, $D(1-D) = S^2$. Por lo tanto, el producto de uno de los valores dados (índice de dificultad) por uno menos ese valor ha de dar otro valor (la varianza). De los valores dados, el único valor que cumple lo anterior es 0.4, pues (0.4)(1-0.4) = 0.24, que es otro valor dado. Por lo tanto, D= 0.4, y la varianza es 0.24. Dado que el índice de homogeneidad corregido suele ser menor que el índice de homogeneidad sin corregir, 0.15 y 0.53 serán los índices de homogeneidad corregidos y sin corregir, respectivamente.

9. a) Debería ser modificado. La alternativa más seleccionada debería ser la correcta.
b) $D_j = 50/350 = 0.14$
c) $S_j^2 = (0.14)(0.86) = 0.12$

10. n = 5

$X_c = A - E/(n-1)$. Luego, $29 = 30 - 4/(n-1)$

11.

Sujeto	Preguntas sin contestar
1	3
2	0
3	0
4	3
5	1

12 a) El ítem 1, pues tiene el menor (0.1) índice de dificultad.
b) No. Pues el ítem 2 ha sido acertado por el 60% de la muestra.
c) Var (item 1) = (0.1)(0.9) = 0.09
Var (item 2) = (0.6)(0.4) = 0.24
Var (item 3) = (1)(0) = 0
El ítem de más varianza es el ítem 2 (0.24).
d) $Xc = X - E/2$. $16 = 20 - E/2$. Luego, E = 8.

13. a) 30/200 = 0.15
b) El 1, pues su índice de dificultad (0.15) es el más bajo. En los otros ítems sus índices de dificultad son: 0.7 (ítem 2), 0.5 (ítem 3), 0.4 (ítem 4) y 0.45 (ítem 5)
c) El 1, pues las opciones incorrectas son más elegidas que la correcta. El 2, pues una alternativa no es elegida. El 3, pues las alternativas incorrectas no tienes frecuencias parecidas.

14. HC produce resultados tanto más diferentes de H cuanto menor sea el número de ítems. Si obtenemos la diferencia entre H y HC en cada tabla obtenemos:

Tabla 1:	0.116	0.173	0.138
Tabla 2:	0.325	0.417	0.378

Luego el test largo, de 12 ítems, es el que tiene diferencias menores: Tabla 1.

TEMA II: MODELO CLÁSICO Y CONCEPTO DE FIABILIDAD

1.- INTRODUCCIÓN

En las Ciencias clásicas (Medicina, Física, Química,...) existen aparatos, con márgenes de error especificados, para medir determinadas características como son la temperatura, la presión sanguínea, el peso, la concentración de determinados elementos químicos, etc. En Psicología no existen instrumentos de medición de la introversión, la actitud hacia el aborto, la aptitud espacial o la habilidad lectora, características que no son susceptibles de una medición directa. Para medir los rasgos psicológicos se han elaborado teorías matemáticas o estadísticas que permiten inferir el nivel de rasgo a partir del rendimiento observado de la persona.

Si elaboramos, por ejemplo, una prueba de atención, una persona obtiene una determinada puntuación X en el test. La cuestión que nos planteamos es si esa X representa una buena manifestación del rasgo auténtico de atención que tiene esta persona. Podemos pensar en las consecuencias que tiene para el psicólogo que un test no proporcione una buena información de los niveles de rasgo. Un psicólogo clínico que utiliza un test de depresión en su labor profesional, debe tener un alto grado de certeza de que las puntuaciones que proporciona el test resultan buenas cuantificaciones de los niveles de depresión de sus pacientes.

La teoría clásica de los tests (a partir de los trabajos iniciales de Spearman) propone un modelo formal, denominado como modelo clásico o modelo lineal clásico, fundamentado en diversos supuestos a partir de los cuales se extraen determinadas consecuencias de aplicabilidad práctica para determinar el grado en que un test informa de los niveles de rasgo.

2.- SUPUESTOS FUNDAMENTALES

El modelo de puntuación verdadera se concreta en un primer supuesto:

(1) $X = V + E$

que indica que la puntuación empírica directa de una persona en un test (X) está compuesta de dos componentes hipotéticos: el nivel de rasgo o puntuación verdadera de la persona (V) y un error de medida (E) que se comete al medir el rasgo con el test. El error de medida se considera una variable aleatoria compuesta por los diferentes factores (propios del sujeto, del test y externos a ambos) que hacen que su puntuación empírica no sea exactamente su nivel de rasgo. Por tanto, el error de medida se establece como la diferencia entre la puntuación empírica y la verdadera:

$E = X - V$

El problema es que E y V resultan en principio desconocidos, si bien podemos obtener información sobre ellos si se plantean determinados supuestos adicionales:

(2) $V = \mathbf{E}[X]$

Definimos la puntuación verdadera de una persona como el valor esperado de las posibles puntuaciones empíricas que puede obtener en el test. Dicho de otro modo, sería el promedio de las puntuaciones empíricas que obtiene la persona en un número elevado de aplicaciones del test.

Del supuesto anterior se desprende que:

$\mathbf{E}[E] = 0$

Asumiendo que X y E son dos variables aleatorias, mientras que la puntuación V de la persona es constante, resulta fácil comprobar la igualdad anterior, puesto que:

$\mathbf{E}[E] = \mathbf{E}[X - V] = \mathbf{E}[X] - \mathbf{E}[V] = \mathbf{E}[X] - V = V - V = 0$

(3) $\rho_{VE} = 0$

Este tercer supuesto nos dice que si en una población conociéramos las puntuaciones V y E de los individuos, la correlación entre ambas variables sería nula. Se asume que puntuaciones verdaderas elevadas (bajas) no tienen porqué tener asociados errores elevados (bajos).

(4) $\rho_{E_j E_k} = 0$

El cuarto supuesto asume que si en una población conociéramos los errores de medida de cada individuo en dos tests diferentes (j y k), dada su condición de aleatoriedad, la correlación entre ambas variables también sería nula.

(5) $\rho_{E_j V_k} = 0$

El quinto supuesto nos indica que si en una población conociéramos las puntuaciones E en un test j y las puntuaciones V en un test k, ambas variables correlacionarían cero.

Ejemplo: Supongamos una población de 5 personas, para las que conocemos sus puntuaciones V, E y X en dos tests diferentes, denominados con los subíndices 1 y 2 (En realidad, sólo podemos conocer las puntuaciones X; las restantes puntuaciones se proponen únicamente por razones didácticas):

V_1	E_1	X_1	V_2	E_2	X_2
12	-2	10	12	0	12
11	0	11	11	-2	9
11	0	11	11	2	13
12	2	14	12	0	12
4	0	4	4	0	4

El lector puede comprobar que se cumplen los supuestos planteados en la página anterior, en la tabla de puntuaciones.

De cualquier forma, insistimos que en la aplicación real de un test sólo se conocen las puntuaciones X de las personas, por lo que los supuestos planteados (por muy lógicos y razonables que sean) no pueden someterse a contrastación empírica, siendo ésta una de las principales limitaciones de la TCT.

3.- CONCEPTO DE FORMAS PARALELAS

Cuando un psicólogo aplica un test a una persona, únicamente conoce su puntuación directa X en la prueba. Lo importante, como venimos indicando, es obtener información de las relaciones entre las X y las V. Un procedimiento sería obtener la correlación entre ambas para un grupo de N personas, pero nos encontramos con el inconveniente de desconocer las auténticas V de las N personas. Sí resulta factible, sin embargo, obtener la correlación entre las puntuaciones empíricas que proporcionan dos formas paralelas de un test, diseñadas ambas para evaluar el mismo rasgo V de los individuos.

Según el modelo clásico, **dos formas paralelas** de un test se definen mediante dos condiciones:

a) Un individuo tiene la misma puntuación V en ambas formas.

b) La varianza de los errores de medida en ambas formas es la misma.

El lector puede comprobar en la tabla de datos expuesta anteriormente que los tests 1 y 2 pueden considerarse formas paralelas, dado que se cumplen en los datos las dos condiciones planteadas. Ahora bien, estamos asumiendo que los datos anteriores se refieren a una población determinada, en la que conocemos las V y los E de los individuos. En la práctica desconocemos esas puntuaciones y, además, disponemos generalmente de datos muestrales y no poblacionales. ¿Cómo determinar entonces si dos formas son o no paralelas? En la tabla anterior podemos constatar que, si dos formas son paralelas, las medias poblacionales de X en ambas son iguales, y también los son las varianzas poblacionales de las puntuaciones X. Según esto, y haciendo uso de los procedimientos empleados en estadística inferencial, si disponemos de datos muestrales podemos realizar los contrastes oportunos para determinar,

con cierta probabilidad, si dos formas son o no paralelas.

Para muestras relacionadas, el contraste sobre diferencia de medias se plantea como:

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

Siendo el estadístico de contraste:

$T = \dfrac{\overline{D}\sqrt{N}}{S_D}$, que sigue la distribución t con N-1 grados de libertad,

donde $\overline{D}$ es la media de las diferencias, N el tamaño de la muestra y S_D la desviación típica insesgada de las diferencias.

El contraste sobre diferencia de varianzas, para muestras relacionadas, se plantea como:

$H_0 : \sigma_1 - \sigma_2 = 0$

$H_1 : \sigma_1 - \sigma_2 \neq 0$

Siendo el estadístico de contraste:

$T = \dfrac{(S_1^2 - S_2^2)\sqrt{N-2}}{2S_1S_2\sqrt{1-r_{12}^2}}$, que sigue la distribución t con N – 2 grados de libertad.

donde r_{12} es la correlación de Pearson entre X_1 y X_2.

Ejemplo: Queremos saber, con probabilidad 0.95, si dos tests (1 y 2) son o no formas paralelas. Aplicamos ambos tests a una muestra de 5 personas y obtienen las siguientes puntuaciones:

X_1	X_2
15	15
10	15
13	20
14	10
18	15

Para el contraste de diferencia de medias, obtenemos un valor T = -0.46, lo que nos lleva a no rechazar H_0, mientras que para el contraste sobre diferencia de varianzas obtenemos un estadístico T = -0.34, que también nos lleva a no rechazar H_0 de igualdad de varianzas

poblacionales. Según esto, podemos decir, con probabilidad 0.95, que ambos tests son formas paralelas.

Los fundamentos de este tipo de contrastes pueden consultarse en el libro de Pardo y San Martín (1998) "Análisis de datos en Psicología II".

4.- SIGNIFICADO DEL COEFICIENTE DE FIABILIDAD

Si dos formas de un test pretenden medir un mismo rasgo, parece razonable esperar que los resultados empíricos de ambas en una población correlacionen de forma elevada. Si esto es así, ambas formas manifiestan un elevado grado de precisión a la hora de reflejar los diversos niveles de rasgo. Si ambas correlacionasen de forma mínima, no podemos fiarnos de que reflejen fidedignamente los niveles de rasgo. Pues bien, definimos inicialmente el **coeficiente de fiabilidad** como la correlación entre los resultados que proporcionan dos formas paralelas. Para datos poblacionales y puntuaciones diferenciales, la expresión de la correlación de Pearson es la siguiente:

$$\rho_{12} = \frac{\Sigma x_1 x_2}{N\sigma_1\sigma_2}$$

Según el primer supuesto del modelo clásico, que se cumple también para escala diferencial, tenemos que x = v + e, con lo que la expresión anterior quedaría como:

$$\rho_{12} = \frac{\Sigma(v_1 + e_1)(v_2 + e_2)}{N\sigma_1\sigma_2}$$

Desarrollando la fórmula anterior, obtenemos que:

$$\rho_{12} = \frac{\Sigma v_1 v_2}{N\sigma_1\sigma_2} + \frac{\Sigma v_1 e_2}{N\sigma_1\sigma_2} + \frac{\Sigma e_1 v_2}{N\sigma_1\sigma_2} + \frac{\Sigma e_1 e_2}{N\sigma_1\sigma_2}$$

Recordando los supuestos del modelo clásico, podemos comprobar que los tres últimos sumandos son iguales a cero, con lo que nos queda la siguiente expresión:

$$\rho_{12} = \frac{\Sigma v_1 v_2}{N\sigma_1\sigma_2}$$

y dado que las puntuaciones v de un mismo individuo en dos formas paralelas las asumimos idénticas, y también son iguales las varianzas poblacionales en ambas formas, la expresión anterior queda como:

$$\rho_{12} = \frac{\Sigma v^2}{N\sigma_1\sigma_2} = \frac{\sigma_v^2}{\sigma_x^2}$$

lo que significa que el coeficiente de fiabilidad es el cociente entre la varianza de las puntuaciones verdaderas y la varianza de las puntuaciones empíricas. Se puede interpretar como la proporción de la varianza empírica que puede atribuirse a la variabilidad de las personas a nivel de rasgo o puntuaciones verdaderas. Nótese además que el coeficiente de fiabilidad puede asumir valores entre 0 y 1, ya que las varianzas siempre serán positivas.

En las siguientes páginas estudiaremos varios procedimientos para calcular el coeficiente de fiabilidad de un test.

5.- FIABILIDAD DE UN TEST DE POR "n" FORMAS PARALELAS

Imaginemos que disponemos de n formas paralelas para medir un rasgo psicológico determinado. Según lo visto, las "n" formas tendrán en la población las mismas varianzas empíricas. Además, las correlaciones entre todos los posibles pares de formas paralelas que podemos establecer serán también iguales, e indicarán la fiabilidad de cualquiera de ellas a la hora de determinar los niveles de rasgo.

Los parámetros de la población en una forma paralela podemos designarlos como σ_x, σ_v, σ_e, y ρ_{xx}. Si unimos las n formas paralelas en un único test, los parámetros de este test alargado podemos expresarlos como σ_{nx}, σ_{nv}, σ_{ne}, y ρ_{nxx}. Vamos a llegar a determinadas expresiones para obtener los parámetros del test alargado conociendo los parámetros de una forma paralela.

a) La **varianza empírica** del test formado por n formas paralelas será:

$$\sigma^2_{nx} = n\,\sigma^2_x + n(n-1)\,\sigma^2_x\,\rho_{xx} = n\,\sigma^2_x\,[1 + (n-1)\,\rho_{xx}]$$

b) La **varianza verdadera** del test formado por n formas paralelas será:

$$\sigma^2_{nv} = n\,\sigma^2_v + n(n-1)\,\sigma^2_v\rho_{vv} = n\,\sigma^2_v\,[1 + (n-1)\,] = n^2\,\sigma^2_v$$

c) La **varianza error** del test formado por n formas paralelas será:

$$\sigma^2_{ne} = n\,\sigma^2_e + n\,(n-1)\,\sigma^2_e\,\rho_{ee} = n\,\sigma^2_e$$

d) A partir de las expresiones anteriores, y recordando que el coeficiente de fiabilidad es el cociente entre la varianza verdadera y la varianza empírica, podemos obtener el **coeficiente de fiabilidad** del test alargado:

$$\rho_{nxx} = \frac{\sigma^2_{nv}}{\sigma^2_{nx}} = \frac{n^2\sigma^2_v}{n^2\sigma^2_x\left(1+(n-1)\rho_{xx}\right)} = \frac{n\rho_{xx}}{1+(n-1)\rho_{xx}}$$

La expresión anterior se conoce como **fórmula general de Spearman-Brown**, y permite obtener el coeficiente de fiabilidad de un test compuesto por n formas paralelas.

Ejemplo: Un test de aptitud para la dirección empresarial está formado por dos formas paralelas de 20 ítems cada una. Aplicados a una población de directivos, se obtiene una correlación de 0.6 entre ambas formas. ¿Cuál será el coeficiente de fiabilidad del test compuesto por la unión de las dos formas paralelas?

$$\rho_{xx} = \frac{n\rho_{xx}}{1+(n-1)\rho_{xx}} = \frac{2(0.6)}{1+(2-1)0.6} = 0.75$$

Comprobamos que el coeficiente de fiabilidad del test alargado (de 40 ítems) es superior al coeficiente de fiabilidad de cualquiera de las formas iniciales de 20 ítems. Esto representa una propiedad interesante del coeficiente de fiabilidad, dado que nos indica que si alargamos un determinado test con formas paralelas, podemos incrementar su fiabilidad.

El razonamiento que hemos expuesto se puede generalizar al caso de que los k ítems que componen un test fueran formas paralelas. En una determinada población, los k ítems de un test serán paralelos si todos tienen la misma media, la misma varianza y la misma fiabilidad. Según la fórmula general de Spearman-Brown, el coeficiente de fiabilidad del test se puede expresar como:

$$\rho_{xx} = \frac{k\rho_{il}}{1+(k-1)\rho_{il}}$$

donde k es el número de ítems del test y ρ_{jl} es la correlación de Pearson entre cualquier par de ítems, que indica la fiabilidad de cada uno de los ítems.

EJERCICIOS

1. Cuatro personas responden a dos tests. Sus puntuaciones en X (conocidas) y en V y E (nunca conocidas, pero supuestamente conocidas en el ejemplo) son las siguientes:

	TEST 1			TEST 2		
	X_1	V_1	E_1	X_2	V_2	E_2
persona 1	3	2	1	0	2	-2
persona 2	2	3	-1	5	3	2
persona 3	4	5	-1	7	5	2
persona 4	7	6	1	4	6	-2

Comprobar qué supuestos de la Teoría Clásica se cumplen y cuales no, en cada test.

2. Un test se aplica a 4 personas. Suponemos conocidas algunas de sus puntuaciones verdaderas y errores. Sabiendo que en los siguientes datos se cumple exactamente la Teoría Clásica, complete las puntuaciones que faltan en la tabla:

	X	V	E
persona 1		5	0
persona 2		7	1
persona 3			0
persona 4			
MEDIA	6		

3. En la aplicación de un test de aptitud numérica, el encargado de controlar el tiempo prolonga 1 minuto el período establecido para resolver las diversas tareas. ¿Cuál es el supuesto de la Teoría Clásica que se vería afectado por tal error, y que por tanto sería difícil de asumir racionalmente?

4. Si dos tests son paralelos, una persona obtendrá la misma puntuación empírica en uno y otro. V ( ) F ( ) Depende ( ). Razone su respuesta.

5. Después de aplicar a 5 personas dos formas de un test de razonamiento analógico, se obtienen los siguientes datos (las desviaciones típicas tienen denominador n-1):

$S_A = 3{,}79$ $S_B = 2{,}83$ $S_D = 1{,}41$ $r_{AB} = 0{,}95$

a) ¿Cuál es la diferencia mínima que deberíamos haber obtenido para considerar, con probabilidad 0.95, que las medias poblacionales son diferentes?

b) Suponiendo que las dos medias no alcanzan esa diferencia mínima, ¿podemos afirmar, con probabilidad 0.95, que ambas formas son paralelas?

6. Si dos formas paralelas de un test se aplican en el mismo momento a un grupo normativo, la correlación entre los resultados de ambas aplicaciones debe ser igual a 1.
V ( ) F ( ). Razone su respuesta.

7. Si la varianza verdadera de un test es el 64 % de su varianza empírica, ¿cuál es su coeficiente de fiabilidad?

8. Complete los valores omitidos en la siguiente tabla, siendo **n** el número de veces que se alarga el test.

	σ^2_x	σ^2_v	σ^2_e	ρ_{xx}	n	ítems
Test original						25
Test alargado	112		16		4	

9. En un test A de 10 ítems la varianza de las puntuaciones verdaderas es 3 y la varianza error es 1. Elaboramos 4 formas paralelas del test A y formamos un nuevo test (test B), resultado de añadir al test A las 4 formas paralelas anteriores. Justifique sus respuestas.

a) El test B tendrá _____ ítems.
b) La varianza de las puntuaciones verdaderas del test B será _____.
c) La varianza de las puntuaciones empíricas obtenidas en el test B será _____.

10. El coeficiente de fiabilidad de un test X de 10 ítems es 0.67. Responda razonadamente.

a) ¿Qué proporción de la varianza de X se debe a las diferencias en los verdaderos niveles de rasgo?
b) Formamos el test doble (de 20 ítems). ¿Qué proporción de la varianza del test doble se debe a los errores de medida?
c) Si correlacionamos las puntuaciones obtenidas entre las dos formas paralelas que forman el test doble, ¿qué correlación esperamos encontrar? ¿Qué proporción de la varianza de las puntuaciones obtenidas en la primera forma depende de las puntuaciones obtenidas en la segunda forma?

11. Diga si las siguientes afirmaciones son ciertas (V) o no (F). No necesita razonar sus respuestas.

a) El índice de homogeneidad de un ítem depende de la relación entre el ítem y las puntuaciones en el test. V( ) F( )
b) Si se aumenta la longitud de un test con ítems paralelos aumentará la varianza error. V( ) F( )
c) Si se aumenta la longitud de un test con ítems paralelos aumentará la varianza verdadera. V( ) F( )
d) Si se aumenta la longitud de un test con ítems paralelos aumentará la varianza empírica. V( ) F( )
e) Según el modelo clásico, los errores de medida NO pueden ser negativos. V( ) F( )
f) En el modelo clásico se asume que las puntuaciones verdaderas y empíricas correlacionan 0 en la población. V( ) F( )

SOLUCIONES

1. X = V + E. Se cumple.
La media de los errores es 0. Se cumple el segundo supuesto.
$\rho_{VE} = 0$. Se cumple el tercer supuesto.
Los errores correlacionan. No se cumple el supuesto 4.
Los errores no correlacionan con las puntuaciones verdaderas ($\rho_{E1\ V2} = \rho_{E2\ V1} = 0$). Se cumple el supuesto 5.

2. Como la media de los errores ha de ser cero, $E_4 = -1$.
Como la correlación entre V y E es cero, tendrá que ser cero su numerador, $\sum(V-\overline{V})(E-\overline{E})$. Es decir, (5-6).(0)+ (7-6).(1)+(V_3-6)(0)+(V_4-6).(-1) = 0, luego, $V_4 = 7$.
Como la media de V ha de coincidir con la media de X, se obtiene $V_3 = 5$
Como X = V + E, $X_1 = 5$; $X_2 = 8$; $X_3 = 5$ y $X_4 = 6$.

3. Si se prolonga el tiempo, cabe suponer que las puntuaciones X de las personas serían superiores a las que les corresponderían con el tiempo bien controlado. En este caso, los errores de medida (E = X - V) serán mayoritariamente positivos, con lo cual se incumple el supuesto de que su media debe ser cero.

4. Depende. El modelo supone que en dos formas paralelas, una misma persona tiene la misma V, pero sus puntuaciones empíricas en una y otra forma por lo general serán diferentes.

5. a) La diferencia mínima es 1.75.
b) El estadístico T para contrastar si las dos varianzas poblacionales son iguales es 1.65, menor que el valor de las tablas (3.182). Aceptamos que son formas paralelas.

6. Falso. No tiene por qué ser 1, ya que las puntuaciones empíricas en una y otra forma no tienen por qué ser las mismas. La correlación entre ambas será un indicador de la fiabilidad de cualquiera de ellas.

7. $r_{xx} = 0.64$

8.

	σ^2_x	σ^2_v	σ^2_e	ρ_{xx}	n	ítems
Test original	10	6	4	0,6	1	25
Test alargado	112	96	16	0,86	4	100

9.

	Items	n	S^2_v	S^2_e
Test A	10	1	3	1
Test B		5		

a) El número de ítems del test B será (5)(10) = 50.
b) $S_{nv}^2 = (n^2)S_v^2 = (25)(3) = 75$
c) 80. Pues $S_{ne}^2 = (n)\,S_e^2 = (5)(1) = 5$ y $S_{nx}^2 = S_{nv}^2 + S_{ne}^2 = 80$.

10. a) El coeficiente de fiabilidad es 0.67. Luego la proporción que piden es 0.67.
b) En el test doble, R = 2(0.67)/(1+0.67) = 0.8. Luego, la proporción que piden es 0.2.
c) La correlación es r_{xx}, que vale 0.67. La proporción pedida es $0.67^2 = 0.45$.

11. a) V
b) V
c) V
d) V
e) F
f) F

TEMA III: FIABILIDAD DEL TEST

1.- INTRODUCCIÓN

Se entiende por fiabilidad el grado de estabilidad, precisión o consistencia que manifiesta el test como instrumento de medición de un rasgo determinado. Si un herrero mide varias veces con una cinta métrica la longitud de una barra de hierro, siempre obtendrá la misma medición, debido a que tanto la cinta métrica como la barra permanecen invariantes. Ahora bien, cuando empleamos un test para medir un rasgo psicosocial determinado, puede ocurrir que ni uno ni otro permanezcan invariantes de una situación a otra; análogamente, sería como disponer de una cinta métrica elástica y de una barra de hierro sometida a diferentes temperaturas (y, por lo tanto, más o menos dilatada). Es labor de la psicometría establecer en cada caso el grado de estabilidad del instrumento de medición.

Hasta el momento, el modelo clásico de puntuación verdadera y el planteamiento de la fiabilidad como correlación entre formas paralelas, se han establecido en términos paramétricos; es decir, suponiendo conocidos los datos de la población de referencia. Lo real es que en la práctica vamos a disponer de datos obtenidos en una muestra o grupo normativo concreto. Esto significa que, de modo directo, únicamente vamos a disponer de las puntuaciones empíricas de dicha muestra, a partir de las cuales podemos obtener los estadísticos que sean oportunos.

Tradicionalmente, la fiabilidad de un test puede entenderse de tres maneras diferentes:

a) Aludiendo a la estabilidad temporal de las medidas que proporciona.

b) Haciendo referencia al grado en que diferentes partes del test miden un rasgo de manera consistente.

c) Enfatizando el grado de equivalencia entre dos formas paralelas.

2.- FIABILIDAD COMO ESTABILIDAD TEMPORAL

Si disponemos de las puntuaciones de N personas en un test y, después de transcurrido un tiempo, volvemos a medir a las mismas personas en el mismo test, cabe suponer que siendo el test altamente fiable, deberíamos obtener una correlación de Pearson elevada entre ambos mediciones. Dicha correlación entre la evaluación test y la evaluación retest (r_{xx}) se denomina **coeficiente de fiabilidad test-retest**, e indicará tanta mayor estabilidad temporal de la prueba cuanto más cercano a uno sea.

Este modo de operar se desprende directamente del modelo lineal clásico, según el cuál se define la fiabilidad como la correlación entre las puntuaciones empíricas en dos formas paralelas, ya que no existe mayor grado de paralelismo entre dos tests que cuando en realidad es uno aplicado dos veces.

Ejemplo: A una muestra de 10 estudiantes de COU se le aplica un cuestionario de hábitos de estudio. Transcurridos dos meses, se vuelve a aplicar el mismo test a las mismas personas bajo las mismas condiciones. Sus puntuaciones directas en las dos aplicaciones fueron las siguientes:

Persona	Test	Restest
1	16	10
2	14	14
3	12	8
4	11	12
5	10	10
6	8	8
7	8	7
8	6	5
9	4	4
10	1	2

Para obtener el coeficiente de fiabilidad test-retest basta con correlacionar los datos de las dos últimas columnas:

$$r_{xx} = 0.87$$

En este caso se obtiene una elevada estabilidad de las puntuaciones. Si los niveles de rasgo (hábitos de estudio) de las personas no han variado a lo largo de los dos meses transcurridos entre las dos aplicaciones, podemos decir que el test proporciona bastantes garantías respecto a la precisión con la que mide, dado que una persona concreta obtiene puntuaciones muy parecidas (o similares) en las dos aplicaciones.

Más concretamente, y haciendo uso del teorema demostrado en el tema anterior, podemos interpretar que el 87 % de la varianza empírica se debe a la variabilidad de las personas a nivel de puntuaciones verdaderas.

Este coeficiente se obtiene, sobre todo, en pruebas cuyo objetivo de medida es un rasgo estable (pruebas de inteligencia general, aptitudes, rasgos de personalidad, etc.) dado que, de lo contrario, no se podría discernir entre la inestabilidad debida al rasgo de la causada por el instrumento de medición. Es aconsejable dejar periodos largos entre la evaluación test y la retest cuando los ítems y las respuestas pueden memorizarse con facilidad; de lo contrario, los sujetos podrían emitir pautas de respuesta similares en las dos aplicaciones del test únicamente por efectos del recuerdo y del deseo de responder de manera congruente, con lo que r_{xx} se incrementaría debido a factores ajenos a la fiabilidad de la prueba. Debe tenerse en cuenta, sin embargo, que cuanto mayor es el intervalo temporal que se deja entre ambas aplicaciones, mayor es la posibilidad de que las puntuaciones de los sujetos oscilen diferencialmente debido a factores de tipo madurativo y, por lo tanto, esto tiene un efecto concreto en el decremento de la correlación entre las puntuaciones del test y del retest.

3.- FIABILIDAD COMO CONSISTENCIA INTERNA

La precisión o fiabilidad de un test se puede entender también como el grado en que diferentes subconjuntos de items miden un rasgo o comportamiento homogéneo; es decir, el grado en que covarían, correlacionan o son consistentes entre sí diferentes partes del cuestionario.

Lo más usual es obtener la consistencia entre dos mitades del test (método de dos mitades) o entre tantas partes como elementos tenga la prueba (consistencia interna).

3.1.- MÉTODO DE DOS MITADES

Este procedimiento consiste en dividir el test en dos mitades equivalentes (normalmente una con los elementos pares y otra con los impares). Para cada sujeto se obtiene la puntuación directa en ambas mitades. Disponemos entonces de dos variables (P e I), cuya correlación de Pearson (r_{PI}) indica su grado de relación.

Si la mitad par e impar fueran entre sí formas paralelas (ya sabemos cómo comprobarlo estadísticamente), la correlación entre ambas sería una medida de la fiabilidad de cada una de ellas. Ahora bien, cuando hemos deducido la fórmula general de Spearman-Brown hemos visto que los tests más largos (con más items) suelen ser más fiables, por lo que r_{PI} estará subestimando el coeficiente de fiabilidad del test total en la medida que P e I son variables extraídas de la mitad de ítems que tiene el test. Para superar este problema, y así obtener el coeficiente de fiabilidad del test completo, debemos aplicar la fórmula de Spearman-Brown, considerando ahora que estamos trabajando con datos muestrales, y haciendo n = 2 ya que el test completo tiene el doble de items que cualquiera de sus mitades:

$$r_{xx} = \frac{2r_{PI}}{1 + r_{PI}}$$

A partir de esta fórmula podemos comprobar que el coeficiente de fiabilidad, entendido como la expresión de la consistencia entre dos mitades, es mayor que la correlación de Pearson entre ambas mitades.

Ejemplo: Supongamos que la siguiente tabla refleja los resultados de una muestra de 10 personas que responden a un cuestionario de 6 ítems valorados de forma dicotómica:

	Ítems								
Sujeto	1	2	3	4	5	6	P	I	Total
1	1	0	1	0	1	0	0	3	3
2	0	1	1	1	0	1	3	1	4
3	0	0	1	0	0	0	0	1	1
4	0	1	1	1	0	0	2	1	3
5	0	0	0	1	0	0	1	0	1
6	1	1	1	1	1	1	3	3	6
7	1	1	1	1	1	1	3	3	6
8	0	1	1	1	0	1	3	1	4
9	0	1	0	0	0	0	1	0	1
10	0	0	0	0	0	0	0	0	0
			Media				1.6	1.3	2.9
			Desviación típica				1.28	1.19	2.02

En este caso se obtiene que $r_{PI} = 0.34$, y por tanto:

$$r_{xx} = \frac{2(0.34)}{1 + 0.34} = 0.51$$

De nuevo el tope de r_{xx} lo tenemos en 1, con lo que podemos decir que las dos mitades del test no son muy consistentes entre sí. Unicamente un 51 % de la varianza de las puntuaciones empíricas se debe a la varianza de las puntuaciones verdaderas. No podríamos afirmar con suficiente certeza que ambas mitades miden con precisión el rasgo de interés.

La razón de dividir el test en la mitad par y la impar es garantizar su equivalencia. Los tests de rendimiento óptimo suelen tener ítems ordenados en dificultad, de tal forma que se comienza a responder los ítems más fáciles hasta llegar a los situados al final del test, que son los más difíciles. Si realizásemos la partición en dos mitades atendiendo a su disposición en la prueba (la primera mitad formada por los primeros n/2 ítems, la segunda por los n/2 ítems últimos) difícilmente podría cumplirse que ambas tuvieran la misma media.

3.2.- COEFICIENTE α DE CRONBACH

En el tema precedente vimos que si los k ítems de un test fueran paralelos, el coeficiente de fiabilidad del test podría obtenerse aplicando la fórmula general de Spearman-Brown:

$$\rho_{xx} = \frac{k\rho_{jl}}{1+(k-1)\rho_{jl}}$$

siendo k el nº de ítems del test y ρ_{jl} la correlación de Pearson entre cualquier par de ítems.

Expresada la fórmula anterior para datos muestrales, quedaría como:

$$r_{xx} = \frac{kr_{jl}}{1+(k-1)r_{jl}}$$

Una fórmula equivalente a la anterior; es decir, que proporciona exactamente el mismo resultado, es la denominada **coeficiente α de Cronbach**:

$$\alpha = \frac{k}{k-1}\left(1-\frac{\sum S_j^2}{S_x^2}\right)$$

donde k es el nº de ítems

$\sum S_j^2$ es la suma de las varianzas de los ítems y S_x^2 es la varianza del test

Dado que las puntuaciones en el test son la suma de las puntuaciones en los ítems, la varianza del test puede expresarse como:

$$S_x^2 = \sum_{j=1}^{k} S_j^2 + 2\sum_{j<l}\mathrm{cov}(j,l)$$

por lo que la expresión inicial puede quedar como:

$$\alpha = \frac{k}{k-1}\left(\frac{2\sum_{j<l}\mathrm{cov}(j,l)}{S_x^2}\right)$$

Esta fórmula reproduce el coeficiente de fiabilidad del test si todos los ítems son paralelos. En la práctica, es muy difícil que esto se produzca pero, sin embargo, tiene sentido su aplicación para establecer el grado en que los diferentes ítems están midiendo una única dimensión o rasgo. Podemos observar en la última expresión que α depende del grado de covariación de los ítems: tendrá un valor alto (cercano a 1) cuando los ítems covaríen fuertemente entre sí; asumirá valores cercanos a cero si los ítems son linealmente independientes (si covarían de forma escasa). Matemáticamente, α puede asumir valores negativos.

Insistimos en que el coeficiente alfa no es un coeficiente de fiabilidad si, como ocurre en la práctica totalidad de los tests, los ítems no son paralelos. Suele considerarse una "estimación por defecto" del coeficiente de fiabilidad, lo que significa que es igual al coeficiente (si los ítems son paralelos) o menor (cuando no lo son). Debe interpretarse como un indicador del grado de covariación entre los ítems, y es aconsejable complementarlo con otras técnicas estadísticas (por ejemplo Análisis Factorial) antes de interpretarlo como una medida de unidimensionalidad.

Ejemplo:

Sujetos	Ítems				X
	1	2	3	4	
1	0	0	0	1	1
2	1	0	0	0	1
3	1	0	0	0	1
4	1	1	1	1	4
5	1	1	0	1	3
6	1	1	0	0	2
Varianzas	0.14	0.25	0.14	0.25	1.33

$$\alpha = \frac{k}{k-1}\left(1-\frac{\sum S_j^2}{S_x^2}\right) = \frac{4}{4-1}\left(1-\frac{0.14+0.25+0.14+0.25}{1.33}\right) = 0.55$$

En este caso, el coeficiente α obtenido representa un valor medio, que nos indica que no existe un elevado grado de covariación entre los ítems. No podemos afirmar con rotundidad que este test mide un rasgo unitario.

El coeficiente α puede obtenerse también entre diferentes grupos de ítems (subtests). En ese caso, k será el número de subtests y ΣS_j^2 la suma de las varianzas de los subtests. Un coeficiente α bajo indicará que los diferentes subtests miden rasgos o constructos diferentes.

4.- FIABILIDAD COMO CORRELACIÓN ENTRE FORMAS PARALELAS

A veces, por razones de índole práctica o investigadora, se diseña un test y una segunda versión del mismo, denominada forma paralela, que intenta evaluar o medir lo mismo que el test original pero con diferentes ítems. Como ya hemos explicado, dos versiones o formas se consideran paralelas si, aplicadas a una misma muestra de personas, obtienen medias y varianzas probabilísticamente similares.

La correlación de Pearson entre las puntuaciones obtenidas en una misma muestra en dos formas paralelas se considera el coeficiente de fiabilidad de cualquiera de ellas, e indicará el grado en que pueden considerarse equivalentes.

Ejemplo:

Sujetos	Forma 1	Forma 2
1	1	4
2	14	12
3	11	13
4	11	9
5	10	12
Medias	9.4	10
Varianzas	19.44	10.8
Varianzas (ins.)	24.3	13.5

$r_{xx} = r_{12} = 0.883$

No es común diseñar una forma paralela de un test para obtener datos sobre su fiabilidad. Cuando se diseñan (tarea por otra parte difícil) es porque van a utilizarse en determinados trabajos que requieren 2 aplicaciones sucesivas de un test que se puede recordar con facilidad.

Por ejemplo, para evaluar la eficacia de ciertos programas cortos de enriquecimiento cognitivo o motivacional, conviene utilizar antes y después del entrenamiento pruebas equivalentes aunque con contenidos diferentes (formas paralelas) para evitar los efectos del recuerdo.

5.- EL ERROR TÍPICO DE MEDIDA

5.1.- CONCEPTO

Asumiendo el postulado fundamental del modelo clásico, que expresa la relación:

$$X = V + E$$

es fácil demostrar que se cumple la siguiente relación para datos muestrales:

$$S_x^2 = S_v^2 + S_e^2$$

A la desviación típica de los errores de medida (S_e) se denomina **error típico de medida**. En cierta manera, el S_e representa también una medida de precisión: cuanto más cercano a cero sea el error típico de medida de un test, eso significará que dicho test proporciona a cada persona una puntuación X cercana a su nivel de rasgo V.

En términos paramétricos, habíamos demostrado en el tema anterior que:

$$\rho_{xx} = \frac{\sigma_v^2}{\sigma_x^2}$$

Para datos muestrales, la expresión anterior queda establecida como:

$$r_{xx} = \frac{S_v^2}{S_x^2} = 1 - \frac{S_e^2}{S_x^2}$$

De donde se deduce que el error típico de medida puede obtenerse a partir de la expresión:

$$S_e = S_x \sqrt{1 - r_{xx}}$$

5.2. APLICACIÓN: CONTRASTE SOBRE PUNTUACIONES VERDADERAS

Un test impreciso puede proporcionar a dos personas puntuaciones empíricas diferentes aunque sus niveles de rasgo sean iguales. Utilizando los procedimientos de las estadística inferencial, podemos contrastar, con cierta probabilidad, si dos puntuaciones empíricas diferentes suponen o no niveles de rasgo distintos.

Para realizar el contraste, para las puntuaciones de dos personas (designadas con los subíndices i y j) planteamos las siguientes hipótesis:

H_0: $V_i = V_j$

H_1: $V_i \neq V_j$

Puede comprobarse que el estadístico de contraste se expresa como:

$$Z = \frac{X_i - X_j}{S_e \sqrt{2}}$$

Si el valor de Z se encuentra dentro de la zona de aceptación, admitiremos, con la probabilidad establecida, que las puntuaciones V de las dos personas son las mismas; de lo contrario, admitiremos que difieren sus niveles de rasgo.

Ejemplo: Un test de Inteligencia general manifiesta en un grupo normativo un coeficiente de fiabilidad de 0.91 y una desviación típica de 16. Dos personas obtienen en el test unas puntuaciones directas de 126 y 120 puntos, respectivamente. ¿Podemos afirmar, con probabilidad 0.95, que ambas personas difieren en sus rasgos intelectuales?.

En este caso, el estadístico será:

$$Z = \frac{126 - 120}{16\sqrt{1 - 0.91}\sqrt{2}} = 0.88$$

Con probabilidad 0.95, la zona de aceptación queda establecida entre los límites Z = -1.96 y Z = 1.96, con lo cual, admitimos con dicha probabilidad que los niveles de rasgo de ambas personas no difieren.

6.- FACTORES QUE AFECTAN A LA FIABILIDAD DE UN TEST

El conocimiento preciso y exhaustivo de los factores que determinan la cuantía del coeficiente de fiabilidad puede ayudarnos en la tarea de diseñar pruebas adecuadas. El tema es relevante en la fase de selección de ítems, para saber cuáles deben seleccionarse dependiendo de los objetivos que se pretenden conseguir. También va a resultar útil para conocer las propiedades y limitaciones que asumimos cuando aplicamos un determinado cuestionario.

Ya hemos aclarado las diferentes versiones que pueden adquirir la fiabilidad de un cuestionario, entendida sobre todo como consistencia o como estabilidad temporal.

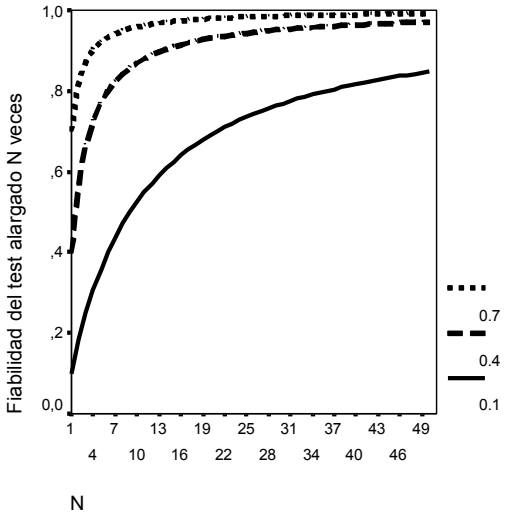
Respecto a la consistencia interna (coeficiente α), basta inspeccionar las dos expresiones formales que hemos proporcionado para comprobar que los elementos que covarían de manera elevada y positiva con los restantes son los que más contribuyen a que α sea elevado. Se puede comprobar, además, que $S^2_x = (\Sigma H_j S_j)^2$, con lo que, si sustituimos la expresión en la fórmula de α, comprobamos también que (entre los elementos de igual variabilidad) los de mayor H_j son los que más contribuyen a incrementar α. Si en la fase de análisis de items tenemos como objetivo elaborar un test con elevada consistencia interna, tenemos que quedarnos con los ítems que manifiestan un mayor índice de homogeneidad.

Además, debe tenerse en cuenta que el coeficiente alfa aumenta cuando incrementamos la longitud del test y que resultaría fácil obtener valores elevados cuando se incluyen ítems redundantes, lo que, evidentemente, no resulta deseable.

En cuanto al coeficiente de fiabilidad (r_{xx}), su cuantía depende en parte de la variabilidad de la muestra donde se obtiene y también de la longitud (número de ítems) del test.

Debemos conocer que un mismo test tiene diferentes r_{xx} en diferentes grupos normativos (muestras de personas donde se obtiene el coeficiente). Más concretamente, un mismo test suele obtener un r_{xx} mayor en un grupo heterogéneo que en otro menos heterogéneo (de menor varianza). Por ejemplo, resulta normal que un test de Inteligencia obtenga un r_{xx} mayor en una muestra de la población general que una muestra de universitarios o en otra de personas con deficiencias cognitivas (estas últimas más homogéneas). La razón es simple: el coeficiente de fiabilidad, obtenido por el método que sea, se fundamenta estadísticamente en una correlación de Pearson que, como es sabido, se incrementa a medida que lo hacen las varianzas de las variables que se correlacionan.

Por otra parte, si los ítems están bien formulados y resultan discriminativos, un test incrementará su r_{xx} a medida que incrementa su longitud (número de ítems), aunque no lo hace de manera lineal. La siguiente gráfica muestra el coeficiente de fiabilidad de un test alargado N veces (N: 1, 2, 3, .. 50), cuando el coeficiente de fiabilidad del test de partida es 0.1, 0.4 y 0.7:



Queremos indicar con la gráfica anterior que el incremento es más significativo cuando el test inicial tiene un número pequeño de ítems y bajo coeficiente de fiabilidad, que cuando el test de partida tiene ya un coeficiente de fiabilidad considerable.

La fórmula general de Spearman-Brown, adaptada ahora a los datos obtenidos en una muestra concreta, permite estimar cuál será el coeficiente de fiabilidad (R_{xx}) de un test que se forma con "n" versiones paralelas de un test inicial que tiene un coeficiente de fiabilidad r_{xx}:

$$R_{xx} = \frac{nr_{xx}}{1+(n-1)r_{xx}}$$

Las n-1 formas añadidas deben ser formas paralelas equivalentes al test inicial; de lo contrario, la fórmula anterior no tiene significado alguno.

Por ejemplo, supongamos que una prueba de atención de 25 ítems obtiene en un grupo normativo un r_{xx}= 0,6. Si se añadieran 75 ítems (tres formas paralelas) al test inicial, el test alargado tendría 100 ítems (4 veces el inicial), y su fiabilidad sería:

$$R_{xx} = \frac{nr_{xx}}{1+(n-1)r_{xx}} = \frac{(4)0.6}{1+(3)0.6} = 0.86$$

Si las 3 formas añadidas fuesen auténticamente paralelas a la original al pasar de 25 a 100 ítems el coeficiente de fiabilidad pasa de 0.6 a 0.86.

Imagínese ahora que el test de atención de 25 ítems tiene un r_{xx} = 0,92. Si se añaden 75 ítems más paralelos, el test alargado tendría de coeficiente de fiabilidad:

$$R_{xx} = \frac{nr_{xx}}{1+(n-1)r_{xx}} = \frac{(4)0.92}{1+(3)0.92} = 0.98$$

En el primer caso, el incremento que se produce al multiplicar por 4 la longitud inicial del test de atención es de 0.26, mientras que en el segundo caso, el incremento es únicamente de 0.06. Esto se debe a que el coeficiente de fiabilidad del test inicial es mayor en el segundo caso que en el primero.

Podemos observar que cuando n = 2 (cuando se duplica la longitud del test original), la fórmula se convierte en la que hemos aplicado para estimar el coeficiente de fiabilidad por el método de las dos mitades. Efectivamente, ahora podemos entender mejor que r_{PI} sería el coeficiente de fiabilidad de un test mitad (con la mitad de los elementos que tiene el test entero) y que el resultado de esa correlación hay que corregirlo, haciendo n = 2 en la fórmula de Spearman-Brown, para obtener el coeficiente de fiabilidad del test completo.

Estas relaciones entre fiabilidad y longitud de un test pueden ayudarnos a estimar el coeficiente de fiabilidad alargando "n" veces o, planteado inversamente, el número de veces que debemos multiplicar la longitud inicial de un test para alcanzar un R_{xx} determinado. En la práctica, puede resultar eficaz diseñar un test inicial corto y estimar cuál debería ser su longitud para alcanzar un coeficiente de fiabilidad determinado, y así comprobar si merece la pena continuar con ítems paralelos o reformar los ya generados. Para ello, si despejamos "n" de la fórmula general de Sperman-Brown, obtenemos la siguiente expresión:

$$n = \frac{R_{xx}(1-r_{xx})}{r_{xx}(1-R_{xx})}$$

Ejemplo: Supongamos que un test inicial de 25 ítems obtiene un coeficiente de fiabilidad de 0.6, considerado bajo para los objetivos que se pretenden conseguir con su aplicación. Una manera de incrementar su precisión es alargarlo con ítems paralelos a los iniciales. Al constructor de la prueba le interesa que el test tenga, al menos, un coeficiente de fiabilidad de 0.86, y se pregunta con cuántos ítems lo conseguiría.

Aplicando la fórmula precedente, obtenemos:

$$n = \frac{0.86(1-0.60)}{0.60(1-0.86)} = 4$$

Esto significa que si multiplicamos por 4 la longitud inicial del test, es decir, con un test de 100 ítems, conseguiremos la precisión deseada. Por tanto, a los 25 ítems que tiene el test inicial habría que añadir 75 ítems paralelos (3 formas) para conseguir la fiabilidad de 0.86.

El lector puede comprobar que este planteamiento es el inverso al del ejemplo precedente, que consideraba los mismos datos, y que por eso es lógico que el resultado de "n" sea 4.

EJERCICIOS

1. Señale el objetivo que se pretende conseguir con cada una de las siguientes actuaciones en la construcción de un cuestionario.

a) Correlacionar las puntuaciones totales en el cuestionario con las mismas puntuaciones obtenidas un mes después.
b) Correlacionar las puntuaciones de la mitad par con las obtenidas en la mitad impar.
c) Valorar todas las covariaciones que se producen entre los diferentes elementos.
d) Correlacionar las puntuaciones totales en el test con las obtenidas por los mismos sujetos en una forma paralela.

2. Señale qué factores pueden afectar al coeficiente de fiabilidad de un test (r_{xx}).

3. Un psicólogo construye una escala de actitudes para evaluar el dogmatismo religioso. La escala consta de 4 ítems, y en cada uno se puede manifestar la opinión según una escala de 7 puntos (del 1 al 7). A continuación se detallan las respuestas de un grupo normativo de 5 personas:

Sujeto	ítem 1	ítem 2	ítem 3	ítem 4
1	1	5	2	7
2	2	3	4	6
3	4	4	3	3
4	5	5	6	7
5	6	7	6	7

Obtenga e interprete el coeficiente α de la prueba.

4. **¿**Cuál sería el coeficiente α de un test formado por 20 ítems completamente independientes entre sí?

5. De un test formado por 40 elementos se sabe que la correlación entre las puntuaciones en los 20 elementos pares y en los 20 elementos impares es igual a 0,5. Obtenga el coeficiente de fiabilidad del test de 40 elementos.

6. Un psicólogo social diseña un test de actitudes hacia la no violencia formado por 4 ítems, en cada uno de los cuales los sujetos pueden manifestar su opinión según una escala de

3 puntos (0, 1 ,2). A continuación se detallan las respuestas dadas por un grupo normativo de 8 personas:

	Sujetos							
	nº 1	nº 2	nº 3	nº 4	nº 5	nº 6	nº 7	nº 8
Ítem nº 1	2	2	2	2	2	2	1	0
Ítem nº 2	2	2	2	2	1	1	0	0
Ítem nº 3	2	1	2	0	0	1	0	0
Ítem nº 4	1	1	0	1	0	0	0	0

Obtenga el coeficiente de fiabilidad de test por el método de dos mitades. Aplique para ello la fórmula de Spearman-Brown .

7. Se aplican dos formas paralelas de un test a un grupo normativo de 10 personas. Sus puntuaciones empíricas directas en ambas formas fueron las siguientes:

Sujetos	1	2	3	4	5	6	7	8	9	10
Forma A	6	3	5	4	4	6	5	5	6	3
Forma B	6	3	4	4	5	6	3	5	6	5

Obtenga el coeficiente de fiabilidad del test según el método de formas paralelas.

8. A continuación se detallan las puntuaciones que un grupo normativo de 4 personas obtiene en un test de personalidad, cuyo coeficiente de fiabilidad fue 0.8:

Sujeto:	1	2	3	4
X_i :	14	6	16	4

a) Calcule la varianza verdadera del test.
b) Calcule el error típico de medida del test.

9. Un psicólogo que trabaja en un centro dedicado a evaluar la rapidez visomotora de los conductores confecciona una pequeña prueba para medir esta habilidad. La prueba consta de 6 elementos que se valoran de forma dicotómica. La tabla siguiente muestra las respuestas que se obtuvieron en un grupo normativo formado por 4 aspirantes a conductores:

a) Obtenga el coeficiente de fiabilidad del test.

b) Obtenga la diferencia mínima que debe producirse entre las puntuaciones de dos conductores en el test para considerar, con probabilidad 0.99, que sus puntuaciones verdaderas son distintas.

	Ítems					
Conductor	1	2	3	4	5	6
1	0	1	1	1	1	1
2	1	1	1	1	0	1
3	0	1	0	1	0	0
4	0	1	1	0	0	0

10. Un test de habilidad verbal de 30 ítems tiene, según el procedimiento de las dos mitades, un coeficiente de fiabilidad de 0.8 y una varianza de 20 puntos.

a) Calcule la correlación entre la mitad par e impar del test.
b) Suponiendo que las dos mitades son auténticamente paralelas, obtenga la varianza de las puntuaciones en la mitad impar del test.
c) Obtenga la covarianza entre las dos mitades.
d) Obtenga la varianza error del test si se le añaden 45 elementos paralelos a los que ya tiene.

11. Sean dos tests de tres ítems. La matriz de correlaciones entre los tres ítems en cada test ha sido:

Correlaciones	Test A	Test B
$item_1$ e $item_2$	0.5	0.3
$item_1$ e $item_3$	0.7	0.4
$item_2$ e $item_3$	0.6	0.4

a) ¿En cual de los tests cabe esperar que sea mayor el coeficiente alfa? Razone su respuesta.
b) ¿En cual de los tests cabe esperar que sea mayor el índice de homogeneidad del ítem 1? Razone su respuesta.

12. Un test está formado por 4 ítems dicotómicos que tienen igual media (0.6). La correlación entre cualesquiera dos de ellos es 1/6. Obtenga el coeficiente alfa del test de 4 ítems.

13. Disponemos de un test inicial, A, de 20 ítems, que tiene un coeficiente de fiabilidad r_{aa}. Multiplicamos su longitud por 2, 3 y 4, siempre con elementos paralelos, y formamos los tests B, C y D, de 40, 60 y 80 ítems, respectivamente. Obtenemos sus coeficientes de fiabilidad: r_{bb}, r_{cc} y r_{dd}. Dado que el test B resulta de añadir 20 ítems al test A; el C, de añadir otros 20 al test B; y el D, de añadir otros 20 al C, ¿cabe esperar que $r_{bb} - r_{aa} = r_{cc} - r_{bb} = r_{dd} - r_{cc}$?

14. Un cuestionario para evaluar el rendimiento en Aritmética está formado por 4 ítems, que se valoran de forma dicotómica (1 el acierto y 0 el fallo). Se aplicó a una muestra de 100 niños. A continuación se detalla alguna información estadística de la mitad par (P), impar (I) y del total del test (X). También aparecen las frecuencias de aciertos (F) de cada uno de los 4 ítems, no habiendo omisiones en ninguno.

Correlaciones :

	P	I	X
P	1		
I	0,.45	1	
X	0,79	0,74	1
Medias	1,50	1,10	2,60
S_j	0,67	0,83	1,14

Ítem	1	2	3	4
F	50	70	60	80

a) Imagínese que aplicamos el test a un niño antes y después de un programa de entrenamiento en aritmética. Diga cuál debe ser la diferencia mínima entre sus dos puntuaciones para considerar, con probabilidad 0.99, que dicho entrenamiento ha tenido eficacia; es decir, para considerar que su nivel de rasgo se ha incrementado.
b) Obtenga e interprete un indicador de la consistencia interna global de la prueba.

15. Un test A tiene 100 ítems y un coeficiente de fiabilidad de 0.5. Un test B tiene el mismo coeficiente de fiabilidad, pero tiene 10 ítems. ¿Significa esto que si a ambos tests añadimos 50 ítems paralelos, los dos tests alargados tendrían la misma fiabilidad?
SI( ) NO( ) Depende ( ). Razone su respuesta.

16. Tenemos un test de 5 ítems con coeficiente de fiabilidad de 0.10.
Aplicando la fórmula $n = R(1-r)/(1-R)r$, para que R = 0.95, n ha de ser 171.

a) ¿Cuántos ítems se han de <u>añadir</u> al test para que su fiabilidad sea 0.95? Realice el cálculo necesario.
b) ¿Puede el valor "n" de la fórmula anterior ser negativo?
SI ( ) NO ( ) DEPENDE ( ). Razone su respuesta.

17. Antonio, Bernardo y Carlos hacen el mismo test y sus puntuaciones son 25, 21 y 28 puntos, respectivamente. Realizado el contraste de igualdad de puntaciones verdaderas entre Antonio y Bernardo, con un nivel de confianza de 0.95, no podemos mantener la hipótesis nula de igualdad de puntuaciones verdaderas.

a) Realizamos el correspondiente contraste, con el mismo nivel de confianza, para comparar las puntuaciones verdaderas de Carlos y Bernardo ¿llegaríamos a la misma decisión que antes sobre sus puntuaciones verdaderas?
b) Realizamos el correspondiente contraste, con el mismo nivel de confianza, para comparar las puntuaciones verdaderas de Carlos y Antonio ¿Mantendríamos la hipótesis nula de igualdad de puntuaciones verdaderas?

SOLUCIONES

1. a) Obtener la fiabilidad test-retest, es decir, la estabilidad temporal de las puntuaciones que proporciona el cuestionario.
b) Obtener la fiabilidad del test mitad. Aplicando la corrección de Spearman-Brown se obtiene la fiabilidad del test completo, según el procedimiento de las dos mitades.
c) Estudiar la consistencia interna del test. Se puede hacer mediante el coeficiente α de Cronbach.
d) Obtener la fiabilidad mediante el método de las formas paralelas.

2. La varianza del grupo normativo y la longitud del test.

3. α = 0.77, que se puede considerar un coeficiente medio-alto. Los 4 ítems covarían entre sí de forma apreciable.

4. α = 0. Si los ítems son independientes, sus covarianzas serán igual a cero.

5. $r_{xx} = 0{,}66$

6. $r_{xx} = 0{,}83$

7. $r_{xx} = 0{,}587$

8. a) $S_v^2 = 20{,}8$
b) $S_e = 2{,}28$

9. a) $r_{xx} = 0{,}778$
b) 2.58 será la diferencia mínima que debe producirse entre dos puntuaciones en el test para considerar, con probabilidad 0,99, que las correspondientes puntuaciones verdaderas son diferentes.

10. a) $r_{P\,I} = 0{,}67$

b) $S_i^2 = 6$
c) $S_{P\,I} = 4$
d) $S^2_{n\,e} = 10$

11. a) El test A. Cuando las correlaciones entre los ítems son más altas, lo serán las covarianzas, y por tanto el coeficiente alfa.
b) El test A. Cuando las correlaciones entre los ítems son altas, también lo serán las correlaciones de cada ítem con el test total (índice de homogeneidad).

12. $S_1^2 = S_2^2 = S_3^2 = S_4^2 = 0{,}24$
$\text{cov}(i,j) = r_{ij} S_i S_j = (1/6)(0.24)^{1/2}(0.24)^{1/2} = (1/6)(0.24)$

$$\alpha = \frac{4}{3}\left(1 - \frac{(4)0.24}{(4)0.24 + 2(6)\frac{1}{6}0.24}\right) = 0.44$$

13. No. A incrementos constantes en longitud, no se producen incrementos constantes en r_{xx}.

14. a) Z = 2.33 en las tablas en contraste unilateral.
$r_{xx} = (2)(0.45)/(1+0.45) = 0.62$
$S_e = 0.70$
Diferencia mínima: $(2.33)(0.70)\sqrt{2} = 2.3$

b) $S_1^2 = 0.25$; $S_2^2 = 0.21$; $S_3^2 = 0.24$; $S_4^2 = 0.16$. $\alpha = \frac{4}{3}\left(1 - \frac{0.86}{1.14^2}\right) = 0.45$.

15. NO. En el primer caso, el test inicial se habrá alargado 1.5 veces para llegar a los 150 ítems del test final. En el segundo, el test inicial ha de alargarse 6 veces, para llegar a los 60 ítems. Partiendo del mismo coeficiente de fiabilidad, normalmente se llega a coeficientes distintos cuando el test se alarga 1.5 y 6 veces.

16. a) Ítems que ha de tener el test = (171)(5) = 855
Ítems a añadir = 855 – 5 = 850.

b) r y R son coeficientes de fiabilidad, por lo que $0 < r, R < 1$. Por lo tanto, en la fórmula, "n" no puede tomar valores negativos. En la fórmula equivalente que se estudia en un tema posterior (validez), entonces sí que "n" puede tomar un valor negativo e indica que el valor de R_{xy} propuesto no es alcanzable alargando el test. En el caso de la fiabilidad, todo valor menor de 1 es alcanzable y "n" da siempre positivo.

17. a) Si se rechaza el contraste de igualdad de puntuaciones verdaderas cuando la diferencia entre las puntuaciones observadas es de 4 puntos, necesariamente se ha de rechazar la igualdad cuando la diferencia es mayor. La diferencia entre Carlos y Bernardo es de 7 puntos.

b) Si se rechaza el contraste de igualdad de puntuaciones verdaderas cuando la diferencia entre las puntuaciones observadas es de 4 puntos, no podemos saber qué sucederá cuando la diferencia sea menor. Se puede aceptar o rechazar la hipótesis nula. La diferencia entre Carlos y Antonio es de 3 puntos. Habrá que hacer el contraste para saberlo.

TEMA IV: VALIDEZ DEL TEST

1.- CONCEPTO DE VALIDEZ

Una cosa es que el test mida de manera precisa o estable (esta cualidad se refiere a su fiabilidad), y otra diferente es la cuestión de qué es lo que auténticamente está evaluando. En el ámbito psicosocial, los diferentes constructos resultan difícilmente operativizables de manera indiscutible, y a veces se producen dudas razonables sobre qué mide un determinado test. Una prueba de inteligencia general tendrá un elevado grado de validez si asigna puntuaciones altas a las personas muy inteligentes, puntuaciones medias a las personas medianamente inteligentes y puntuaciones bajas a las personas de poca inteligencia. Un cuestionario para evaluar el nivel de autoestima tendrá un elevado nivel de validez si se demuestra que mide de forma exhaustiva todos los componentes en que puede manifestarse la autoestima.

La validación es un proceso continuo, que incluye procedimientos diferentes para comprobar si el cuestionario mide realmente lo que dice medir. Dicho de otro modo, tiene que ver con el tipo de conclusiones o inferencias que pueden realizarse a partir de las puntuaciones obtenidas en el test. Las inferencias pueden ser de muy diverso tipo: ¿qué rasgo estamos midiendo realmente? ¿Qué podemos predecir sobre el comportamiento de un sujeto que obtiene una determinada puntuación en el test? ¿Qué consecuencias de diverso tipo tiene esa puntuación, en contextos de evaluación o selección?

Aunque cada vez se tiende más a concebir la validez como un proceso unitario que tiene como objetivo aportar pruebas sobre las inferencias que podemos realizar con un test, tradicionalmente se han diferenciado varios procedimientos de validación, alguno de los cuales incluye varios métodos diferentes de comprobación. Los fundamentales procedimientos son denominados como validez de contenido, de constructo y referida al criterio.

2.- VALIDEZ DE CONTENIDO

Sobre todo en pruebas de rendimiento (por ejemplo, pruebas de inteligencia, de aptitudes, etc...) y en pruebas de conocimientos (cuestionarios para evaluar el rendimiento en una materia escolar o en una especialidad temática concreta), tiene sentido justificar que el conjunto de items que forman el test conforman una muestra representativa del universo de contenidos que interesa evaluar. Un test de conocimientos de Química en 8º de EGB, por ejemplo, debería incluir cuestiones representativas de los diferentes núcleos de contenidos que oficialmente deben impartirse en ese nivel de estudios. Sería una prueba poco válida si incluye demasiadas cuestiones de unos temas y muy pocas de otros.

Para justificar, aunque sólo sea racionalmente, que un test posee validez de contenido, debe quedar bien definido el universo o dominio conductual de referencia: especificar claramente cuáles son los contenidos de Química que debe conocer un alumno de 4º de ESO, cuáles son los componentes que interesa considerar en un cuestionario de cultura general, qué tipo de conocimientos y destrezas son las pertinentes para medir el nivel básico de inglés, etc. En

definitiva, nos referimos a explicitar claramente los objetivos de la evaluación y la importancia que se quiere dar a cada uno, lo que determinará la cantidad de cuestiones a incluir referidas a cada uno de esos objetivos. En definitiva, la validez de contenido es un tema particular del de muestreo: si deseamos realizar inferencias sobre el rendimiento de las personas en una población de contenidos determinada, el test debe incluir una muestra representativa de dichos contenidos.

El proceso de validación de contenido es eminentemente lógico, si bien pueden utilizarse jueces expertos en el tema para valorar la congruencia entre los diversos items y los diversos objetivos. Existen procedimientos cuantitativos diversos para que cada experto valore el grado en que un ítem sirve para evaluar el objetivo al que corresponde. El procedimiento cuantitativo más sencillo sería el siguiente:

- Especificar los diversos objetivos (v.gr. áreas diferentes de contenidos) que se pretenden evaluar.
- Elaborar varios ítems para cada objetivo.
- Seleccionar una muestra de expertos en el contenido del test.
- Pedirles que, según su opinión, asignen cada ítem al objetivo que pretende medir.
- Seleccionar los ítems en los que los expertos manifiestan mayor acuerdo en sus clasificaciones.

Muy en relación con la validez de contenido se encuentra lo que se ha dado en llamar "**validez aparente**", que se refiere al grado en que un test da la impresión a los evaluandos de que mide lo que se pretende. En situaciones aplicadas, es importante que las personas perciban que los ítems del test tienen que ver con la finalidad que se persigue con el procedo de evaluación.

3.- VALIDEZ DE CONSTRUCTO

Un constructo es un concepto elaborado por los teóricos de la Psicología para explicar el comportamiento humano. Inteligencia fluida, extroversión, autoconcepto, asertividad, motivación intrínseca... son constructos que forman parte de teorías psicológicas y que precisan de indicadores observables para su estudio. En muchas ocasiones, estos indicadores son los ítems de un test, y debe comprobarse empíricamente que resultan adecuados para reflejar el constructo de referencia

3.1.- ESTRATEGIAS PARA LA VALIDEZ DE CONSTRUCTO

La validez de constructo incluye la planificación y ejecución de determinados estudios de investigación orientados a comprobar empíricamente que un test mide realmente el constructo o rasgo que pretendemos.

Aunque los métodos a emplear son sin duda variados, así como la técnicas estadísticas para analizar los datos, podemos encontrar un común denominador a todos ellos, que se sintetiza en las siguientes fases:

1.- **Formular hipótesis relevantes** (extraídas de deducciones teóricas o del sentido común) en las que aparezca el constructo que pretendemos evaluar con el test. En definitiva, una hipótesis de trabajo consiste en poner en relación dos o más variables. Pues bien, una de esas variables ha ser el constructo que pretendemos medir con el test.

2.- **Efectuar en la práctica mediciones** oportunas de las variables o constructos involucrados en las hipótesis. La medición del constructo de interés se realizará con la prueba diseñada a tal efecto, que es la que pretendemos validar.

3.- **Determinar si se verifican o no las hipótesis** planteadas. En el caso de que así sea, queda confirmado mediante una investigación que el test mide el constructo de interés ya que, de lo contrario, no habría razones lógicas para que se cumplieran las hipótesis formuladas. Si las hipótesis no se confirman no significa en principio que el test no es válido, ya que puede ser debido a que las hipótesis no estaban planteadas de manera adecuada, lo cual exigiría una revisión de la teoría subyacente.

Imaginemos, por ejemplo, que un investigador está interesado en validar una prueba de motivación intrínseca-extrínseca que ha construido. Desde la teoría motivacional de partida se puede deducir que las personas motivadas intrínsecamente (por el mero placer que les supone la ejecución de determinadas tareas) deberían rendir mejor en actividades escolares que las personas motivadas por razones extrínsecas (deseos de alcanzar determinada nota o determinado refuerzo externo). Para validar su prueba, el investigador tiene que demostrar empíricamente que mide auténticamente el constructo motivacional que se pretende, y podría proceder de la siguiente manera:

a) Aplicar el test a un grupo amplio de alumnos del nivel escolar apropiado.

b) Recoger información de cada alumno sobre su nivel intelectual, su calificación académica media en el último curso y las horas que dedica al estudio.

c) Formar dos grupos diferentes (A y B), de tal manera que ambos tengan un mismo nivel intelectual medio y que ocupen un número similar de horas en el estudio, pero que el grupo A tenga niveles altos de motivación intrínseca y el B niveles altos de motivación extrínseca.

d) Comparar el rendimiento académico de los dos grupos. Si la hipótesis de partida fuera cierta, el grupo A debería rendir significativamente más que el grupo B, con lo cuál se aportaría información sobre la validez del test. Desde luego, si el test no midiera motivación, sería improbable que se verificase la hipótesis de trabajo.

Pueden ser muy variados los métodos a seguir que, cumpliendo el proceso de ejecución planteado anteriormente, sirvan para poner a prueba la validez de constructo de un test. En cada caso habrá que seguir el que más convenga para contrastar las hipótesis de partida, pero algunos métodos suelen ser más frecuentes. Entre ellos destacamos:

- Obtener las relaciones entre las puntuaciones en el test y en otras variables que deberían relacionarse con el constructo de interés. Si el modelo teórico está bien fundamentado,

debe establecer relaciones entre el constructo de interés y otros diferentes, y por tanto debe ser posible establecer diseños de investigación para contrastar las previsiones teóricas. Por ejemplo, Moltó (1988) predice (y comprueba) que la escala de susceptibilidad al castigo (que mide el grado de evitación de situaciones reales aversivas) debe proporcionar puntuaciones relacionadas directamente con neuroticismo e inversamente con estabilidad emocional.

- Evaluar mediante el test a grupos que se supone deben ser diferentes en el constructo, para comprobar si realmente es así. Resulta un enfoque eminentemente diferencial: si el test es válido, debería reflejar las diferencias entre grupos que se predicen desde la teoría psicológica. Por ejemplo, si un test de inteligencia general para edades infantiles es válido, debería reflejar el mayor rendimiento de los niños de más edad.

- Utilizar una estrategia experimental para comprobar si el test resulta sensible para detectar los efectos previsibles debidos a la manipulación o selección de los niveles en una o más variables independientes. El ejemplo expuesto anteriormente sobre motivación y rendimiento puede servir para entender esta estrategia.

- Aplicar la técnica multivariada del Análisis Factorial (exploratorio o confirmatorio) sobre la matriz de correlaciones entre items, para descubrir estadísticamente las variables o dimensiones subyacentes (factores) a la covariación entre los elementos.

3.2.- VALIDEZ DE CONSTRUCTO FACTORIAL

Este último método, denominado **validez de constructo factorial**, requiere alguna precisión que puede ser pertinente por fundamentarse en una técnica estadística relativamente sofisticada y, sobre todo, porque su utilización práctica es muy extensa.

El **análisis factorial** es una técnica estadística multivariante que sirve para estudiar las dimensiones que subyacen a las relaciones entre varias variables. Normalmente toma como datos de partida la matriz de correlaciones entre las n variables que interesa analizar. Como información final, proporciona una matriz de tamaño $n \times p$, denominada matriz factorial rotada. Esta matriz contiene las *saturaciones* de cada variable en cada una de las "p" dimensiones extraídas, y que son las correlaciones de Pearson entre cada variable y cada dimensión.

El análisis factorial se realiza con dos objetivos 1) determinar cual es el número de dimensiones o factores que mide un test y descubrir cual es el significado de cada una; 2) obtener la puntuación de cada sujeto en cada dimensión. Normalmente, el número de dimensiones que mide un test es mucho menor que el de ítems. Para descubrir su significado y darles sentido es necesario fijarse en las variables que saturan de forma elevada en cada dimensión. Cuando el investigador se enfrenta con la tarea de dar significado a una dimensión, debe realizar un proceso inferencial para encontrar el nexo de unión entre las variables que manifiestan correlaciones elevadas en la dimensión. Además, los diferentes factores (dimensiones) extraídos no tienen la misma importancia. Cada uno explica una determinada cantidad de la varianza total de los ítems, que se expresa porcentualmente, y que indica la importancia de esa dimensión para dar cuenta de la covariación entre las variables. Si un factor explica un porcentaje elevado de la

varianza total, eso es síntoma de que las saturaciones de las variables en dicho factor son altas, lo que significa que es una dimensión importante a la hora de describir las relaciones entre las variables originales.

3.2.1.- EJEMPLO DE ANÁLISIS FACTORIAL

Un psicólogo ha elaborado una prueba de cinco ítems para evaluar la actitud hacia las nuevas tecnologías por parte de las personas mayores. Los ítems, que se responden en una escala de siete categorías ordenadas (desde 1: "muy en desacuerdo" hasta 7: "muy de acuerdo"), son los siguientes:

ítem 1: El uso de teléfonos móviles puede hacerme la vida más fácil.
ítem 2: Los aparatos modernos son demasiado caros.
ítem 3: Me gustaría tener una agenda electrónica.
ítem 4: El coste de las llamadas desde un móvil es razonable.
ítem 5: Gracias a internet podemos resolver muchos problemas.

Los 5 ítems se aplicaron a una muestra de 200 personas. La matriz de correlaciones entre ellos se sometió a un análisis factorial, obteniéndose los siguientes resultados. Esta matriz contiene las saturaciones, es decir, la correlación de cada ítem con cada uno de los factores que mide el test:

Ítem	Factor I	Factor II
1	0.845	-0.126
2	-0.201	0.803
3	0.672	0.012
4	0.052	-0.615
5	0.713	-0.143
% de varianza total explicada	34%	21%

Hay dos factores fundamentales que explican las relaciones entre los 5 items. Supongamos que se tipifican las puntuaciones en los ítems; la varianza total sería cinco, que es la suma de la varianza de cada ítem. El factor I explica un 34% de la varianza total, el factor II explica un 21% de la varianza total. Con los dos factores se explica el 55% de la varianza de los ítems.

En el factor I obtienen saturaciones altas los items 1, 3 y 5, que indican si la persona considera que las nuevas tecnologías pueden ser útiles para mejorar su calidad de vida. El ítem 2 tiene una saturación negativa (aunque baja) porque posiblemente manifiesta una actitud contraria hacia las nuevas tecnologías. Por tanto, el factor I puede denominarse "**Actitud positiva hacia las nuevas tecnologías como medio para mejorar la calidad de vida**".

En el factor II obtienen saturaciones elevadas (en valor absoluto) los ítems 2 y 4, mientras que el resto de saturaciones son cercanas a cero. El hecho de que el ítem 2 tenga una saturación positiva y el 4 negativa significa que las personas con puntuación alta en el factor II tienden a estar de acuerdo con el ítem 2 y en desacuerdo con el 4. Este segundo factor podría etiquetarse "**Sensibilidad hacia el gasto que supone utilizar las nuevas tecnologías**".

Vemos, pues, que las relaciones de covariación entre los ítems podemos explicarlas con dos dimensiones que resultan bastante claras de identificar. Como el lector puede suponer, las cosas no son tan evidentes en la realidad; el investigador debe decidir cuántos factores están presentes en los datos y, sobre todo, debe asignar un significado a cada factor, lo que normalmente no es tan sencillo como en este ejemplo. Lo cierto es que la aplicación del análisis factorial aporta información sobre las dimensiones que estamos midiendo con un determinado cuestionario, es decir, proporciona información sobre la validez de la prueba.

En las siguientes secciones se describe más detalladamente como se obtiene e interpreta la estructura factorial que subyace a las respuestas a los ítems de un test.

3.2.2.- EL MODELO FACTORIAL

El análisis factorial se basa en un modelo que es una extensión del utilizado en teoría clásica de tests. A modo de ejemplo, consideremos los siguientes seis ítems de una escala de *Cordialidad* dirigida a población infantil:

1. Me comporto de manera honesta y correcta con los demás.
2. Trato a mis compañeros afectuosamente.
3. Si un compañero tiene dificultades, le ayudo.
4. Confío en los demás.
5. Pienso que otras personas son buenas y honradas.
6. Dejo que los demás usen mis cosas.

Estos ítems se aplicaron a una muestra de 564 chicos y chicas de entre 11 y 14 años. La matriz de correlaciones obtenida en esta muestra fue:

$$\begin{bmatrix} 1 & & & & & \\ r_{21} & 1 & & & & \\ r_{31} & r_{32} & 1 & & & \\ r_{41} & r_{42} & r_{43} & 1 & & \\ r_{51} & r_{52} & r_{53} & r_{54} & 1 & \\ r_{61} & r_{62} & r_{63} & r_{64} & r_{65} & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ 0{,}459 & 1 & & & & \\ 0{,}313 & 0{,}384 & 1 & & & \\ 0{,}246 & 0{,}285 & 0{,}240 & 1 & & \\ 0{,}171 & 0{,}274 & 0{,}227 & 0{,}448 & 1 & \\ 0{,}150 & 0{,}281 & 0{,}266 & 0{,}286 & 0{,}239 & 1 \end{bmatrix}$$

Observe que unos ítems correlacionan más entre sí que otros. En realidad, el patrón de correlaciones nos informa de cuántas dimensiones subyacen a las respuestas en esos ítems. A continuación se verá que, utilizando el análisis factorial, seremos capaces de extraer muchísima información sobre los ítems a partir de esa matriz de correlaciones.

3.2.2.1. El modelo factorial con un único factor

Puede plantearse que los seis ítems miden una misma característica, la "*cordialidad*". Aplicando el modelo de la teoría clásica de tests a las puntuaciones de los ítems, se obtienen las ecuaciones:

$$X_1 = V + E_1$$
$$X_2 = V + E_2$$
$$X_3 = V + E_3$$
$$X_4 = V + E_4$$
$$X_5 = V + E_5$$
$$X_6 = V + E_6$$

Esto significa que todos los ítems miden la misma característica (la *cordialidad*), representada por *V* en el modelo. Además, hay un error de medida que puede ser distinto para cada ítem (E_i). Es posible plantearse que no todos miden igual de bien la *cordialidad*. Por esta razón se definen los parámetros λ_i, denominados **saturaciones**, que indican la relación de cada ítem con la característica o factor que miden todos ellos. Cuanto mayor sea λ_i, mejor indicador de la *cordialidad* será la puntuación en el ítem. Si a la *cordialidad* se la denomina *F* en lugar de *V*, se obtiene el modelo de un factor:

$$X_1 = \lambda_1 F + E_1$$
$$X_2 = \lambda_2 F + E_2$$
$$X_3 = \lambda_3 F + E_3$$
$$X_4 = \lambda_4 F + E_4$$
$$X_5 = \lambda_5 F + E_5$$
$$X_6 = \lambda_6 F + E_6$$

Las saturaciones se calculan a partir de la matriz de correlaciones entre los ítems. Los cálculos exigen la aplicación de cálculos de álgebra matricial y se realizan mediante ordenador. Sin embargo, la lógica es fácil de ejemplificar. Asumamos que las variables $X_1, X_2, X_3, X_4, X_5, X_6$ y *F* están en puntuaciones típicas. *Si el modelo unidimensional fuera cierto*, la correlación esperada entre X_1 y X_2 (que denominaremos r_{12}^*) sería:

$$r_{12}^* = \frac{\Sigma X_1 X_2}{N}$$

Lo que se simplifica a:

$$r_{12}^* = \frac{\Sigma(\lambda_1 F + E_1)(\lambda_2 F + E_2)}{N}$$

Asumiendo que los errores no correlacionan entre sí ni con la puntuación en el factor (como en la Teoría Clásica), obtenemos que la correlación esperada según el modelo sería igual al producto de los pesos de los 2 ítems en el factor:

$$r_{12}^* = \frac{\lambda_1 \lambda_2 \Sigma F^2}{N} = \lambda_1 \lambda_2$$

Para entender la última simplificación, debe recordar que $\frac{\Sigma F^2}{N} - \overline{F}^2$ es la varianza de las puntuaciones *F;* al estar las puntuaciones *F* en puntuaciones típicas su varianza es 1. Por tanto, sabiendo que $\overline{F} = 0$:

$$\frac{\Sigma F^2}{N} = 1$$

Si calculáramos cuales son los valores esperados de las correlaciones según el modelo unidimensional, a las que denominamos correlaciones *reproducidas* (r^*), obtendríamos la siguiente matriz:

$$\begin{bmatrix} 1 & & & & & \\ r_{21}^* & 1 & & & & \\ r_{31}^* & r_{32}^* & 1 & & & \\ r_{41}^* & r_{42}^* & r_{43}^* & 1 & & \\ r_{51}^* & r_{52}^* & r_{53}^* & r_{54}^* & 1 & \\ r_{61}^* & r_{62}^* & r_{63}^* & r_{64}^* & r_{65}^* & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ \lambda_2\lambda_1 & 1 & & & & \\ \lambda_3\lambda_1 & \lambda_3\lambda_2 & 1 & & & \\ \lambda_4\lambda_1 & \lambda_4\lambda_2 & \lambda_4\lambda_3 & 1 & & \\ \lambda_5\lambda_1 & \lambda_5\lambda_2 & \lambda_5\lambda_3 & \lambda_5\lambda_4 & 1 & \\ \lambda_6\lambda_1 & \lambda_6\lambda_2 & \lambda_6\lambda_3 & \lambda_6\lambda_4 & \lambda_6\lambda_5 & 1 \end{bmatrix}$$

En términos generales, los programas de análisis factorial buscan aquellos valores de λ (λ_1, λ_2, λ_3, λ_4, λ_5 y λ_6) que hacen que las correlaciones esperadas según el modelo (r_{21}^*, r_{31}^*, r_{41}^*, r_{51}^*, …) se parezcan lo más posible a las correlaciones observadas (0,459, 0,313, 0,246, 0,171,…). En el ejemplo, a partir de la matriz de correlaciones mencionada anteriormente, se llega al siguiente modelo de un factor:

$$X_1 = 0{,}540F + E_1$$
$$X_2 = 0{,}671F + E_2$$
$$X_3 = 0{,}542F + E_3$$
$$X_4 = 0{,}529F + E_4$$
$$X_5 = 0{,}483F + E_5$$
$$X_6 = 0{,}437F + E_6$$

Lo cual significa que el factor tiene una relación más fuerte con el ítem 2 que con los demás, aunque todas las saturaciones son elevadas. En el caso de un factor, las saturaciones resultan ser iguales a las correlaciones de cada ítem con el factor. Pueden tomar valores positivos o negativos. Si la saturación es cero, o próxima a cero, no existe relación entre el ítem y el factor. Saturaciones extremas, en cualquier dirección, significan que la relación es fuerte. Generalmente, en los programas informáticos, las saturaciones se disponen en una matriz que se denomina *matriz factorial:*

Matriz factorial[a]

	Factor
	1
x1	.540
x2	.671
x3	.542
x4	.529
x5	.483
x6	.437

Método de extracción: Máxima verosimilitud.

a. 1 factores extraídos. Requeridas 4 iteraciones.

Las correlaciones esperadas según el modelo serían:

$$\begin{bmatrix} 1 & & & & & \\ r^*_{21} & 1 & & & & \\ r^*_{31} & r^*_{32} & 1 & & & \\ r^*_{41} & r^*_{42} & r^*_{43} & 1 & & \\ r^*_{51} & r^*_{52} & r^*_{53} & r^*_{54} & 1 & \\ r^*_{61} & r^*_{62} & r^*_{63} & r^*_{64} & r^*_{65} & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ 0{,}362 & 1 & & & & \\ 0{,}293 & 0{,}363 & 1 & & & \\ 0{,}286 & 0{,}355 & 0{,}287 & 1 & & \\ 0{,}261 & 0{,}324 & 0{,}262 & 0{,}256 & 1 & \\ 0{,}236 & 0{,}293 & 0{,}237 & 0{,}231 & 0{,}211 & 1 \end{bmatrix}$$

Según el modelo de un factor los dos ítems que más deberían correlacionar son los ítems 2 y 3 puesto que son los que más correlacionan con ese factor. Las correlaciones *reproducidas* se parecen a las correlaciones observadas en nuestra muestra, *pero no son iguales*. La diferencia entre una correlación observada y una *reproducida* se llama *residual*:

$$\begin{bmatrix} 1 & & & & & \\ r_{21} - r^*_{21} & 1 & & & & \\ r_{31} - r^*_{31} & r_{32} - r^*_{32} & 1 & & & \\ r_{41} - r^*_{41} & r_{42} - r^*_{42} & r_{43} - r^*_{43} & 1 & & \\ r_{51} - r^*_{51} & r_{52} - r^*_{52} & r_{53} - r^*_{53} & r_{54} - r^*_{54} & 1 & \\ r_{61} - r^*_{61} & r_{62} - r^*_{62} & r_{63} - r^*_{63} & r_{64} - r^*_{64} & r_{65} - r^*_{65} & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ 0{,}096 & 1 & & & & \\ 0{,}020 & 0{,}021 & 1 & & & \\ -0{,}040 & -0{,}070 & -0{,}046 & 1 & & \\ -0{,}090 & -0{,}050 & -0{,}035 & 0{,}192 & 1 & \\ -0{,}086 & -0{,}012 & 0{,}029 & 0{,}055 & 0{,}028 & 1 \end{bmatrix}$$

Por ejemplo, el residual para la correlación entre los ítems 1 y 3 ($r_{31} - r^*_{31}$) es 0,020.

A partir del modelo de un factor, y teniendo en cuenta las propiedades de las combinaciones lineales de variables, la varianza de un ítem puede calcularse como una función de su saturación en el factor, de la varianza del factor y de la varianza del error. Por ejemplo, sabiendo que:

$$X_1 = 0{,}540F + E_1$$

la varianza de X_1 ($\sigma^2_{X_1}$) puede calcularse como:

$$\sigma^2_{X_1} = 0{,}540^2 \sigma^2_F + \psi^2_1$$

donde σ^2_F y ψ^2_1 representan la varianza de F y la varianza de E_1. Al estimar el modelo factorial a partir de la matriz de correlaciones, se está asumiendo implícitamente que los ítems y el factor vienen expresados en puntuaciones típicas. Esto significa que las varianzas del factor y del ítem son 1 ($\sigma^2_F = 1$, $\sigma^2_{X_1} = 1$); Por tanto, la varianza del ítem (1) se descompone del modo siguiente:

$$1 = 0{,}540^2 (1) + \psi^2_1$$

Como se puede ver, una parte de la varianza del ítem depende de su saturación en el factor común. A esa parte se la denomina **comunalidad** y se la representa por el símbolo h^2_i. El resto de la varianza del ítem depende de la varianza del error (ψ^2_1). A esa parte se la denomina **unicidad.** Simbólicamente**,**

$$1 = h^2_1 + \psi^2_1$$

La **comunalidad** de un ítem indica la cantidad de su varianza explicada por el factor. En el modelo de un factor, la comunalidad de un ítem se obtienen elevando la saturación de ese ítem en el factor al cuadrado. En el ejemplo, las comunalidades son $h^2_1 = 0{,}292$ (que es 0.540^2), $h^2_2 = 0{,}450$ (que es 0.671^2), $h^2_3 = 0{,}294$, $h^2_4 = 0{,}280$, $h^2_5 = 0{,}234$ y $h^2_6 = 0{,}191$.

La varianza de los errores se denomina **unicidad**, y se simboliza, como ya hemos mencionado, mediante ψ^2_i. La unicidad de un ítem indica cuanta varianza del mismo no depende del factor, es decir, es varianza específica del ítem que no se relaciona con lo que los ítems miden en

conjunto. Las unicidades se calculan $\psi_i^2 = 1 - h_i^2$. En el ejemplo, las unicidades son $\psi_1^2 = 0{,}708$, $\psi_2^2 = 0{,}550$, $\psi_3^2 = 0{,}706$, $\psi_4^2 = 0{,}720$, $\psi_5^2 = 0{,}766$ y $\psi_6^2 = 0{,}809$.

La suma de las comunalidades es la varianza de los ítems explicada por el factor. En nuestro ejemplo, esta suma es 1,740. Como la varianza total de los ítems es 6, la proporción de varianza explicada por el factor es 1,740/6 = 0,290; es decir, el 29%. La suma de las unicidades es la varianza de los ítems no explicada por el factor, en este caso el 71% del total.

3.2.2.2. El modelo factorial con más de un factor

Es posible formular modelos factoriales en los que cada ítem mida más de una característica simultáneamente. Supongamos que se hipotetiza que el cuestionario mide dos factores, denominados F_1 y F_2. Entonces, las saturaciones se denominan λ_{ij} (siendo i el ítem y j el factor), y el modelo de dos factores es:

$$
\begin{aligned}
X_1 &= \lambda_{11}F_1 + \lambda_{12}F_2 + E_1 \\
X_2 &= \lambda_{21}F_1 + \lambda_{22}F_2 + E_2 \\
X_3 &= \lambda_{31}F_1 + \lambda_{32}F_2 + E_3 \\
X_4 &= \lambda_{41}F_1 + \lambda_{42}F_2 + E_4 \\
X_5 &= \lambda_{51}F_1 + \lambda_{52}F_2 + E_5 \\
X_6 &= \lambda_{61}F_1 + \lambda_{62}F_2 + E_6
\end{aligned}
$$

Las correlaciones reproducidas según el modelo de 2 factores independientes serían:

$$
\begin{bmatrix}
1 & & & & & \\
r_{21}^* & 1 & & & & \\
r_{31}^* & r_{32}^* & 1 & & & \\
r_{41}^* & r_{42}^* & r_{43}^* & 1 & & \\
r_{51}^* & r_{52}^* & r_{53}^* & r_{54}^* & 1 & \\
r_{61}^* & r_{62}^* & r_{63}^* & r_{64}^* & r_{65}^* & 1
\end{bmatrix} =
$$

$$
\begin{bmatrix}
1 & & & & & \\
\lambda_{21}\lambda_{11} + \lambda_{22}\lambda_{12} & 1 & & & & \\
\lambda_{31}\lambda_{11} + \lambda_{32}\lambda_{12} & \lambda_{31}\lambda_{21} + \lambda_{32}\lambda_{22} & 1 & & & \\
\lambda_{41}\lambda_{11} + \lambda_{42}\lambda_{12} & \lambda_{41}\lambda_{21} + \lambda_{42}\lambda_{22} & \lambda_{41}\lambda_{31} + \lambda_{42}\lambda_{32} & 1 & & \\
\lambda_{51}\lambda_{11} + \lambda_{52}\lambda_{12} & \lambda_{51}\lambda_{21} + \lambda_{52}\lambda_{22} & \lambda_{51}\lambda_{31} + \lambda_{52}\lambda_{32} & \lambda_{51}\lambda_{41} + \lambda_{52}\lambda_{42} & 1 & \\
\lambda_{61}\lambda_{11} + \lambda_{62}\lambda_{12} & \lambda_{61}\lambda_{21} + \lambda_{62}\lambda_{22} & \lambda_{61}\lambda_{31} + \lambda_{62}\lambda_{32} & \lambda_{61}\lambda_{41} + \lambda_{62}\lambda_{42} & \lambda_{61}\lambda_{51} + \lambda_{62}\lambda_{52} & 1
\end{bmatrix}
$$

De forma genérica, para un número p el número de factores *independientes*:

$$r_{ij}^* = \sum_{x=1}^{p} \lambda_{ix}\lambda_{jx}$$

De nuevo, el programa busca aquellos valores de λ que hacen que las correlaciones esperadas según el modelo (r_{21}^*, r_{31}^*, r_{41}^*, r_{51}^*, …) se parezcan lo más posible a las correlaciones observadas (0,459, 0,313, 0,246, 0,171,…). Al estimar las saturaciones a partir de la matriz de correlaciones se obtiene el resultado:

$$
\begin{aligned}
X_1 &= 0{,}576F_1 + 0{,}151F_2 + E_1 \\
X_2 &= 0{,}729F_1 + 0{,}222F_2 + E_2 \\
X_3 &= 0{,}469F_1 + 0{,}239F_2 + E_3 \\
X_4 &= 0{,}199F_1 + 0{,}678F_2 + E_4 \\
X_5 &= 0{,}174F_1 + 0{,}606F_2 + E_5 \\
X_6 &= 0{,}270F_1 + 0{,}340F_2 + E_6
\end{aligned}
$$

Los programas nos informarán de esas saturaciones, mediante una matriz denominada *matriz de factores rotados:*

Matriz de factores rotados[a]

	Factor	
	1	2
x1	.576	.151
x2	.729	.222
x3	.469	.239
x4	.199	.678
x5	.174	.606
x6	.270	.340

Método de extracción: Máxima verosimilitud.
Método de rotación: Normalización Varimax con Kaiser.
a. La rotación ha convergido en 3 iteraciones.

Puede verse que los ítems 1, 2 y 3 tienen una correlación más fuerte con el factor I que con el factor II, mientras que ocurre lo contrario para los ítems 4, 5 y 6. Viendo las saturaciones y el contenido de los ítems, puede suponerse que el factor I significa "*Trato a los demás*", mientras que el factor II podría indicar "*Confianza en los demás*".

Al haber dos factores *independientes*, las comunalidades se calculan mediante $h_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2$. Las comunalidades de los 6 ítems serían 0,354, 0,581, 0,277, 0,500, 0,397 y 0,188. Las

unicidades se calculan del mismo modo que en el modelo de un factor ($\psi_i^2 = 1 - h_i^2$), y son 0,646, 0,419, 0,723, 0,500, 0,603 y 0,812. Las correlaciones reproducidas según el modelo serían:

$$\begin{bmatrix} 1 & & & & & \\ r_{21}^* & 1 & & & & \\ r_{31}^* & r_{32}^* & 1 & & & \\ r_{41}^* & r_{42}^* & r_{43}^* & 1 & & \\ r_{51}^* & r_{52}^* & r_{53}^* & r_{54}^* & 1 & \\ r_{61}^* & r_{62}^* & r_{63}^* & r_{64}^* & r_{65}^* & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ 0{,}453 & 1 & & & & \\ 0{,}306 & 0{,}395 & 1 & & & \\ 0{,}217 & 0{,}295 & 0{,}255 & 1 & & \\ 0{,}192 & 0{,}262 & 0{,}227 & 0{,}445 & 1 & \\ 0{,}207 & 0{,}272 & 0{,}208 & 0{,}284 & 0{,}253 & 1 \end{bmatrix}$$

La matriz de *residuales* sería:

$$\begin{bmatrix} 1 & & & & & \\ r_{21} - r_{21}^* & 1 & & & & \\ r_{31} - r_{31}^* & r_{32} - r_{32}^* & 1 & & & \\ r_{41} - r_{41}^* & r_{42} - r_{42}^* & r_{43} - r_{43}^* & 1 & & \\ r_{51} - r_{51}^* & r_{52} - r_{52}^* & r_{53} - r_{53}^* & r_{54} - r_{54}^* & 1 & \\ r_{61} - r_{61}^* & r_{62} - r_{62}^* & r_{63} - r_{63}^* & r_{64} - r_{64}^* & r_{65} - r_{65}^* & 1 \end{bmatrix} = \begin{bmatrix} 1 & & & & & \\ 0{,}005 & 1 & & & & \\ 0{,}006 & -0{,}011 & 1 & & & \\ 0{,}029 & -0{,}010 & -0{,}015 & 1 & & \\ -0{,}021 & 0{,}012 & 0{,}000 & 0{,}003 & 1 & \\ -0{,}056 & 0{,}009 & 0{,}058 & 0{,}002 & -0{,}013 & 1 \end{bmatrix}$$

Indicando que con el modelo de dos factores las diferencias entre las correlaciones *reproducidas* y las correlaciones observadas es muy pequeña.

Además, también es posible calcular **la varianza explicada por cada factor**. Esta varianza es la suma de las saturaciones en ese factor al cuadrado; es decir, la varianza explicada por el factor *x* es $\lambda_{1x}^2 + \lambda_{2x}^2 + \lambda_{3x}^2 + \lambda_{4x}^2 + \lambda_{5x}^2 + \lambda_{6x}^2$. Genéricamente, siendo *n* el número de variables:

$$\sum_{i=1}^{n} \lambda_{ix}^2$$

En el ejemplo, la varianza explicada por cada factor es 1,226 y 1,072, que, al ser la varianza total 6, representa un porcentaje del 20 % y del 18 % respectivamente. Por tanto, el porcentaje de varianza explicada por el modelo de dos factores es del 38 % aproximadamente.

3.2.3.- SELECCIÓN DEL NÚMERO DE FACTORES

Se ha aplicado el modelo de un factor y el modelo de dos factores a los mismos datos. Las correlaciones *reproducidas* según el modelo de un factor se parecían bastante a las correlaciones observadas en la muestra. Al extraer dos factores, se parecían más aún. Por tanto, parece que las correlaciones se *ajustan* un poco mejor al modelo de dos factores. ¿Justifica eso *extraer* dos factores? ¿o deberíamos analizar lo que ocurre cuando extraemos tres factores?

En primer lugar, al añadir un nuevo factor al modelo, aumenta la varianza explicada. En el ejemplo, la varianza explicada ha pasado del 29% al 38% al añadir un segundo factor. Al extraer un factor más, el porcentaje de varianza se incrementaría. Sin embargo, este incremento de la varianza no siempre justifica que se añadan nuevos factores. Lo esperable, y deseable, es que el número de dimensiones sea mucho menor que el de ítems. Esto significaría que una misma característica está siendo medida por muchos ítems a la vez, lo que redundaría en una mejor precisión de la medida.

El número de factores no se decide de forma caprichosa y se han propuesto varios métodos para determinar cuál es el número óptimo de factores que deben retenerse. Esta decisión es muy importante, pues el primer objetivo del análisis factorial es determinar cuantas dimensiones está midiendo un test, es decir, cuantos factores deben incluirse en la solución factorial.

El procedimiento más riguroso estadísticamente se basa en la utilización de criterios de bondad de ajuste, mediante el estadístico X^2 de Pearson. En los apartados anteriores, obtuvimos los residuales en el modelo de dos factores (diferencias entre las correlaciones reproducidas y las observadas en la muestra). Se recordará que esos residuales eran distintos de cero. Parecería que el modelo de dos factores no reproducía de forma perfecta las correlaciones observadas en la muestra. Sin embargo, esto no supone un problema por sí mismo. En realidad, lo correcto es plantearse si podemos mantener que los residuales son cero en la *población. S*abemos que por mera variación muestral es improbable que los residuales sean cero en la muestra incluso si lo son en la población. Así pues mediante el estadístico X^2 contrastamos si podemos mantener la Hipótesis nula de que *todos* los *residuales* son cero en la población.

Veamos un ejemplo. Primero, estimamos varios modelos con distinto número de factores. El modelo más simple (con menos factores) cuyo ajuste sea satisfactorio es el que se utiliza para interpretar los resultados y obtener conclusiones. La siguiente tabla muestra los valores del estadístico X^2 para los modelos de uno y dos factores, los grados de libertad (*gl*) y el nivel crítico (*p*).

Factores	X^2	*gl*	*p*	*RMSEA* *(intervalo de confianza del 90%)*
1	82,213	9	0,000	0,121 (0,098-0,145)
2	10,776	4	0,029	0,055 (0,016-0,096)

Utilizando un nivel de significación α = 0,01, puede concluirse que el modelo de un factor no ajusta a los datos. En cambio, con ese nivel de significación, podemos mantener que el modelo de dos factores se ajusta a los datos. Si utilizamos α = 0,05 diríamos que ninguno de los 2 modelos se ajustaría a los datos.

Un problema asociado al estadístico X^2 es que hace que nuestra decisión sobre el número de factores a retener dependa mucho del tamaño de la muestra. Si la muestra es suficientemente grande, residuales muy pequeños pueden resultar significativos y se tenderá a extraer un número de factores mayor que el necesario desde el punto de vista práctico. En ese caso, el criterio

basado en el estadístico X^2 es excesivamente exigente y poco realista, pues ningún teórico espera que un modelo factorial ajuste de forma *perfecta* a los datos. Por el contrario, si la muestra es pequeña, residuales de valor elevado pueden no resultar estadísticamente significativos y extraeremos un número de factores menor que el necesario.

Algunos autores han propuesto utilizar indicadores de ajuste que nos permitan evaluar el grado de discrepancia entre las correlaciones reproducidas y las correlaciones observadas en la muestra. El RMSEA (*Root Mean Square Error of Approximation*) es uno de esos indicadores. Valores por debajo de 0,05 indican *buen* ajuste del modelo a los datos, valores entre 0,05 y 0,08 indican ajuste *aceptable,* valores entre 0,08 y 0,10 indican ajuste *marginalmente aceptable* y valores por encima de 0,10 indican *mal* ajuste. Si bien no hay que tomar esa clasificación como las "*Tablas de la Ley",* estas guías pueden servir de orientación para tomar una decisión sobre el número de factores a retener. En nuestro ejemplo, el modelo de un factor muestra *mal* ajuste. Siguiendo este criterio podríamos mantener el modelo de dos factores (RMSEA = 0,055) que muestra un ajuste *aceptable*. Además puede observarse que el modelo de un factor y el modelo de dos factores difieren claramente en el RMSEA (ver los intervalos de confianza para el RMSEA).

Para tomar una decisión sobre el número de factores a retener, ayuda observar los residuales. En nuestro caso, se observa que los mayores residuales para el modelo de un factor se encuentran para las correlaciones entre los ítems 4 y 5 (0,192). Ese residual positivo nos indica que esos dos ítems correlacionan entre *sí más de lo que se esperaría si el modelo de un factor fuera cierto*. Naturalmente, cuando extraemos el segundo factor esos dos ítems pesan en él. Su contenido es muy similar (ítem 4: *Confío en los demás*; ítem 5: *Pienso que otras personas son buenas y honradas*).

Existen otros procedimientos de extracción más sencillos pero también muy criticados como la *regla de Kaiser* (*regla K1)* ó el *Scree test.* Una descripción de los métodos de extracción y reglas disponibles en el paquete SPSS puede encontrarse en Pardo y Ruiz[2] (2002). Otros métodos como *el método de análisis paralelo o la regla MAP de Vellicer* también han sido recomendados.

3.2.4.- ROTACIONES

Cuando se estima un modelo factorial las saturaciones no siempre son fácilmente interpretables, en el sentido de que pueden no indicar con claridad qué es lo que están midiendo los factores. Para interpretar la solución, los ítems se agrupan en factores, y el significado de éstos se infiere analizando qué tienen en común los ítems que se agrupan en un mismo factor. Esto no siempre es fácil de descubrir, por ejemplo, si los ítems agrupados en un mismo factor son muy heterogéneos y no tienen un contenido común. Además, hemos visto que los ítems pueden tener saturaciones relativamente altas en más de un factor, lo que significa que miden más de una característica y hace más difícil descubrir su significado.

[2] Pardo, A. y Ruíz, M.A. (2002). SPSS 11. Guía para el análisis de datos. Madrid: Mc Graw Hill.

En nuestro ejemplo, inicialmente obtendríamos la siguiente matriz factorial (*no rotada*) de saturaciones:

Matriz factorial[a]

	Factor	
	1	2
x1	.537	-.257
x2	.700	-.303
x3	.512	-.121
x4	.591	.388
x5	.525	.349
x6	.425	.084

Método de extracción: Máxima verosimilitud.

a. 2 factores extraídos. Requeridas 5 iteraciones.

Según está estructura, el primer factor sería un factor general en el que pesan todos los ítems. En el segundo factor, los pesos mayores son para los ítems 4 y 5 (positivos) y para el ítem 2 (negativo). En principio, esta estructura es difícil de interpretar.

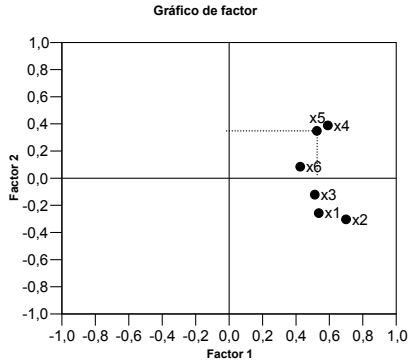
Para facilitar la interpretación se aplica a las saturaciones un proceso denominado rotación, por el cual se transforman las saturaciones en otras más sencillas de interpretar. Con la rotación se intenta que la solución factorial se aproxime a la denominada **estructura simple**. Una estructura simple implica que: a.) en cada factor pesan alto un conjunto de variables (y pesan bajo o cero las restantes variables). b.) los conjuntos de ítems definiendo cada factor no deben solaparse demasiado. c.) cada variable pesa solo en un conjunto pequeño de factores (y pesa bajo o cero en el resto de los factores). Por ejemplo, si la solución factorial hubiera sido:

	F_1	F_2
X_1	0,9	0,0
X_2	0,0	0,7
X_3	0,8	0,0
X_4	0,0	0,6
X_5	0,7	0,0
X_6	0,0	0,8

Esta solución sería más fácilmente interpretable que la que hemos obtenido porque no hay ítems que saturen en ambos factores. En la realidad, mediante las rotaciones nunca se encuentra una estructura simple sino una solución lo más parecida posible a la estructura simple. Veremos a continuación, cuál es la estructura más simple que podemos obtener en nuestro ejemplo.

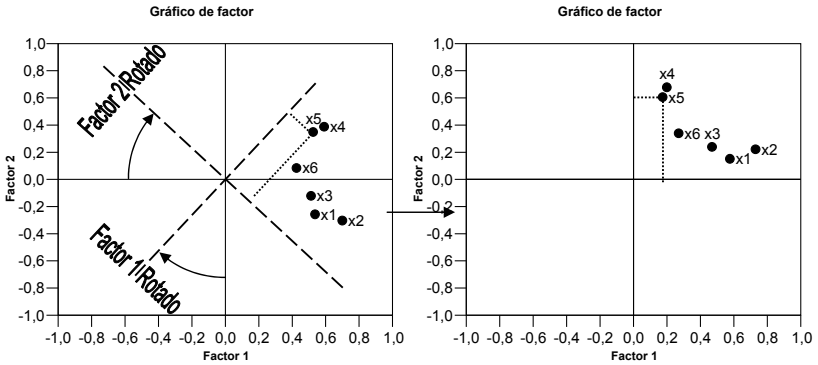
3.2.4.1. ROTACIÓN ORTOGONAL

Las saturaciones encontradas en la solución factorial pueden representarse en un espacio con tantas dimensiones como factores. En el ejemplo, el resultado sería el siguiente, donde los ejes son los factores y cada punto representa las saturaciones de una de las variables (por ejemplo, el ítem 5 tenía saturaciones 0,525 y 0,349 en los factores 1 y 2, respectivamente).



La rotación ortogonal consiste en cambiar (girar) los ejes de referencia (los factores) un cierto número de grados. Observa lo que ocurre cuando cambiamos los ejes de la siguiente forma:

ANTES DE LA ROTACIÓN DESPUÉS DE LA ROTACIÓN



Las posiciones relativas de las variables en el espacio factorial *no cambian.* Al cambiar los ejes, por ejemplo, las nuevas saturaciones del ítem 5 serían 0,174 y 0,606. La matriz factorial rotada sería:

Matriz de factores rotados[a]

	Factor	
	1	2
x1	.576	.151
x2	.729	.222
x3	.469	.239
x4	.199	.678
x5	.174	.606
x6	.270	.340

Método de extracción: Máxima verosimilitud.
Método de rotación: Normalización Varimax con Kaiser.
a. La rotación ha convergido en 3 iteraciones.

Observe que la matriz factorial rotada es más fácil de interpretar que la matriz factorial no rotada. Sin embargo, al *rotar* no cambian las *comunalidades* (ni las *unicidades*) y tampoco las correlaciones reproducidas según el modelo. Por ejemplo:

	Matriz factorial (no rotada)	Matriz de factores rotados
h_1^2	$0{,}537^2 + (-0{,}257)^2 = 0{,}354$	$0{,}576^2 + 0{,}151^2 = 0{,}354$
r_{12}^*	0.537*0,700+(-0,257*-0,303)=0,453	0,576*0,729+0,151*0,222=0,453

Sí cambia el porcentaje de varianza explicada por cada factor (pero no el total de varianza explicada por los dos factores en su conjunto):

% de Varianza explicado por	Matriz factorial (no rotada)	Matriz de factores rotados
Factor 1	30,755	20,428
Factor 2	7,539	17,866
% Total	**38,294**	**38,294**

Este es un resultado general de la rotación ortogonal: la varianza explicada por cada factor cambia después de la rotación, pero no la varianza explicada en total.

El tipo de rotación que se ha utilizado en este apartado es la denominada VARIMAX. Consiste en mover los ejes de referencia, manteniéndolos ortogonales entre sí, para que las saturaciones sean lo más diferentes posible entre si, con lo que se intenta que tomen valores extremos o valores próximos a cero y se eviten los valores intermedios. **Con la rotación VARIMAX los factores son independientes** (correlacionan 0 entre sí).

3.2.4.2. ROTACIÓN OBLÍCUA

La rotación oblicua es más compleja que la ortogonal porque permite que cada factor se rote un número de grados diferente. En el ejemplo, aplicando la denominada rotación oblicua (el método OBLIMIN) se llega a la solución:

Matriz de configuración[a]

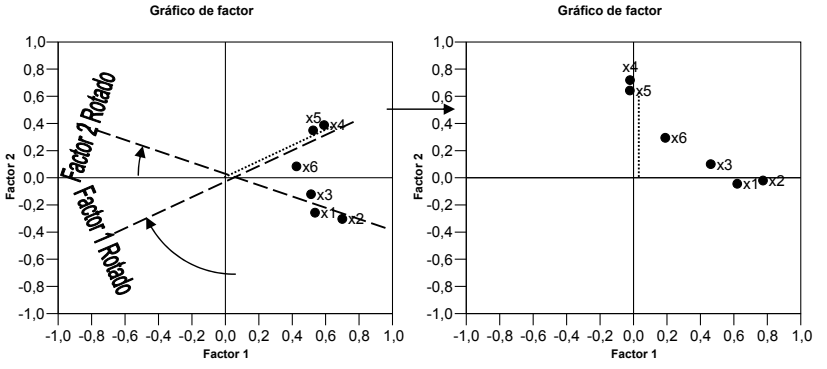
	Factor	
	1	2
x1	.620	-.044
x2	.774	-.020
x3	.462	.100
x4	-.021	.719
x5	-.022	.643
x6	.190	.294

Método de extracción: Máxima verosimilitud.
Metodo de rotación: Normalización Oblimin con Kaiser.
a. La rotación ha convergido en 6 iteraciones.

La siguiente figura muestra las saturaciones obtenidas tras la rotación oblicua. A diferencia de lo que sucedía en los ejemplos anteriores, los ejes de coordenadas (factores) no son perpendiculares. Estadísticamente, esto significa que las puntuaciones en los dos factores están correlacionadas. En el ejemplo, la correlación es de 0,586.

ANTES DE LA ROTACIÓN

DESPUÉS DE LA ROTACIÓN
(Factor 1 y 2 correlacionados)



Al cambiar los ejes, por ejemplo, las nuevas saturaciones del ítem 5 serían -0,022 y 0,643. Puede verse que la solución rotada es más sencilla porque los ítems tienen saturaciones altas en un factor y bajas en el otro. Viendo las saturaciones y el contenido de los ítems, puede suponerse que el factor I significa "*Trato a los demás*", mientras que el factor II podría indicar "*Confianza en los demás*". Como hay una correlación positiva entre los dos factores, los sujetos que tienden a ser cordiales y afectuosos en el trato también suelen confiar en los demás.

De nuevo, al *rotar* no cambian las *comunalidades* (ni las *unicidades*) y tampoco las correlaciones reproducidas según el modelo (aunque con esta rotación, el cálculo de las comunalidades y de las correlaciones reproducidas es más complejo).

La solución obtenida tras la rotación oblicua tiene tres características específicas que deben tenerse en cuenta: 1) las saturaciones ya no son las correlaciones de los ítems con los factores, 2) no es posible determinar la varianza explicada por cada factor, y 3) los factores pueden estar correlacionados. Estas características no se dan en la solución inicial del análisis factorial ni en la obtenida tras la rotación ortogonal.

En resumen, en la práctica el análisis factorial se aplica en dos pasos. En primer lugar se obtiene la solución inicial, lo que permite evaluar la bondad de ajuste del modelo y determinar el número de factores. En segundo lugar se realiza una rotación, ortogonal u oblicua, según los propósitos del investigador. La solución rotada sirve para interpretar el sentido de los factores. Si se realiza la rotación ortogonal, es posible calcular las comunalidades, unicidades y la varianza explicada por cada factor. Si se realiza la rotación oblicua, se obtiene la correlación entre factores y unas saturaciones más sencillas de interpretar.

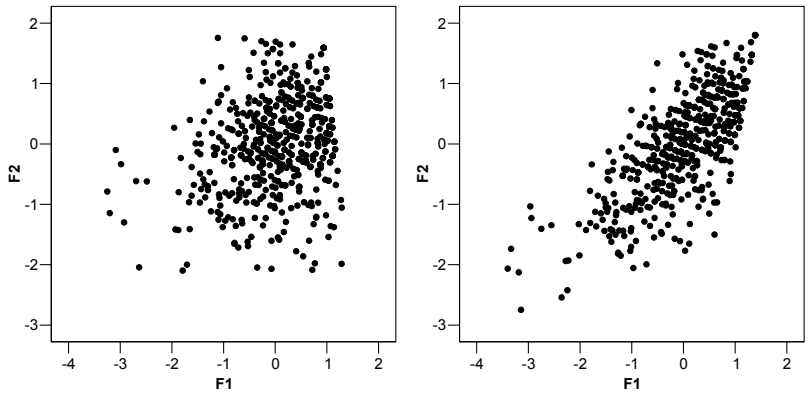
3.2.5.- PUNTUACIONES FACTORIALES

Obtenida una solución factorial definitiva, es posible calcular la puntuación de los sujetos en cada uno de los factores. De este modo, en lugar de obtener una puntuación única para cada sujeto en el test, se obtiene la puntuación en cada uno de los factores que se están midiendo.

La siguiente tabla muestra las respuestas de los cinco primeros sujetos, sus puntuaciones factoriales correspondientes a la rotación factorial y la oblicua. Al haber concluido que el test mide dos factores sería incorrecto utilizar la puntuación en el test como el resultado de cada sujeto. En su lugar, habría que utilizar las dos puntuaciones factoriales correspondientes a la rotación que finalmente se decida aplicar.

							Oblicua		*Ortogonal*	
Sujeto	X_1	X_2	X_3	X_4	X_5	X_6	F_1	F_2	F_1	F_2
1	3	2	3	5	2	1	-1,77	-0,34	-1,95	0,27
2	4	3	3	3	3	3	-0,88	-0,47	-0,85	-0,22
3	3	2	1	1	2	4	-2,28	-1,94	-1,93	-1,42
4	5	3	2	2	2	2	-1,01	-1,40	-0,65	-1,26
5	2	4	1	4	3	1	-1,37	-0,43	-1,44	0,01

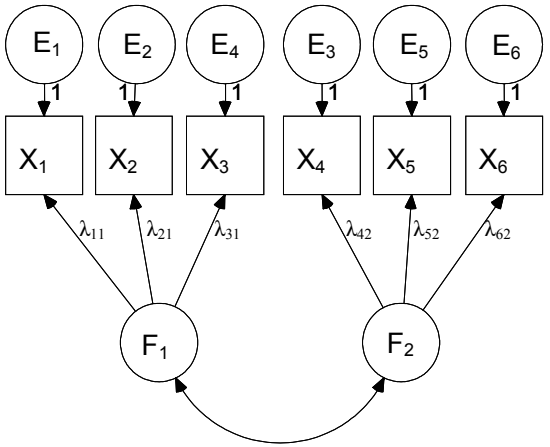
En la siguiente figura aparecen los diagramas de dispersión de las puntuaciones factoriales de los 564 sujetos del ejemplo. El diagrama izquierdo corresponde a la rotación ortogonal y el derecho a la oblicua. El diagrama derecho muestra que existe una relación entre las puntuaciones en ambos factores debida a la correlación existente entre los factores. Esto no sucede así en el izquierdo.



3.2.6. EL ANÁLISIS FACTORIAL CONFIRMATORIO

Cómo ya hemos mencionado anteriormente, el **análisis factorial** sirve para estudiar las dimensiones que subyacen a las relaciones entre varias variables. En realidad hay dos estrategias distintos de análisis factorial: **exploratorio** y **confirmatorio**. Hasta ahora hemos visto como se realiza el primer tipo. En un análisis factorial exploratorio, el investigador no tiene una idea exacta de cuantos factores subyacen a las relaciones entre variables ni tampoco de qué variables tienen un peso alto en cada factor. En el análisis factorial confirmatorio, por el contrario, el investigador plantea *hipótesis definidas a priori* sobre cuál es el número de factores y cómo pesan las variables en ellos. A medida que se acumulan estudios dentro de un campo de conocimiento, los investigadores prefieren utilizar técnicas confirmatorias frente a exploratorias. Aún así, la teoría que subyace a ambos tipos de análisis factorial es la misma.

Imagínese que en la prueba de *Cordialidad*, basándose en estudios previos, establece que los ítems 1, 2 y 3 deben conformar un factor de "trato a los demás" mientras que los ítems 4, 5 y 6 deben conformar un factor de "confianza en los demás". Esto se podría representar de la siguiente manera:



Las variables se representan mediante cuadrados si son observables (como x_1 ó x_5) o mediante círculos si son variables no observables (como los factores F_1 y F_2, o los errores). Las relaciones entre variables se representan mediante líneas. Una línea curva bidireccional conectando dos variables indica que ambas están correlacionadas (en la figura, los Factores F_1 y F_2 están correlacionados). Las flechas rectas direccionales indican que hay una relación direccional entre las 2 variables conectadas (por ejemplo, x_1 recibe líneas de F_1 y del error para representar que está influenciada por ambas variables). La figura anterior se traduciría al siguiente conjunto de ecuaciones:

$$X_1 = \lambda_{11}F_1 + E_1$$
$$X_2 = \lambda_{21}F_1 + E_2$$
$$X_3 = \lambda_{31}F_1 + E_3$$
$$X_4 = \lambda_{42}F_2 + E_4$$
$$X_5 = \lambda_{52}F_2 + E_5$$
$$X_6 = \lambda_{62}F_2 + E_6$$

Observa que no se contemplan efectos de F_2 a X_1 o X_2, ni tampoco de F_1 a X_4 o X_5. Esto es importante. La principal diferencia entre el modelo de análisis factorial exploratorio y el modelo de análisis factorial confirmatorio es que en el último se han fijado algunos pesos al valor *cero* (por ello no se representan en la figura). De esta manera, la solución factorial obtenida en el análisis factorial confirmatorio es *única* y la solución que obtenemos es la que debemos interpretar (no es necesaria la *rotación).*

Para nuestro ejemplo, obtendríamos la siguiente matriz factorial:

	Factor 1	Factor 2
X1	0,595	0
X2	0,745	0
X3	0,536	0
X4	0	0,690
X5	0	0,616
X6	0	0,440

En principio, la hipótesis que se planteaba el investigador no parece desencaminada, ya que los ítems pesan en su factor. La correlación que el programa estima entre los dos factores es 0,617. La siguiente tabla muestra los valores del estadístico X^2 para ese modelo, los grados de libertad (*gl*), el nivel crítico (*p*), el RMSEA y su intervalo de confianza.

X^2	*gl*	*p*	*RMSEA* *(intervalo de confianza del 90%)*
22,874	8	0,004	0,057 (0,031-0,086)

Utilizando un nivel de significación $\alpha = 0,05$, puede concluirse que el modelo *no* se ajusta *perfectamente* a los datos. Sin embargo el RMSEA (0,057) muestra que el modelo tiene un ajuste *aceptable* (está entre 0.05 y 0.08).

3.2.7.- VALIDEZ CONVERGENTE-DISCRIMINANTE. MATRICES MULTIRASGO-MULTIMÉTODO

Muy en relación con la validez factorial se encuentra también la **validez convergente-discriminante**, la cual se evalúa mediante las **matrices multirrasgo-multimétodo**. El sentido de estas definiciones es el siguiente:

> **Validez convergente**. Si dos tests miden un mismo rasgo, la correlación entre ellos debe ser alta.
>
> **Validez discriminante**. Si dos tests miden rasgos diferentes, las correlación entre ellos debe ser baja, o al menos menor que con otro test que mida el mismo rasgo.

Supongamos que desean medirse los rasgos factor *g*, razonamiento espacial y neuroticismo. Cada uno de los rasgos se evalúa mediante dos baterías de tests: A y B. Estos tests se aplican a una muestra de sujetos y se obtiene la siguiente matriz de correlaciones multirasgo-multimétodo, denominada así porque se evalúan varios rasgos utilizando varios métodos.

	A			B		
A	Factor *g*	Espacial	Neuroticismo	Factor *g*	Espacial	Neuroticismo
Factor *g*	0,87					
Espacial	0,61	0,81				
Neurocicismo	0,25	0,31	0,73			
B						
Factor *g*	0,65	0,41	0,09	0,81		
Espacial	0,35	0,50	0,11	0,33	0,78	
Neurocicismo	-0,05	0,08	0,62	0,19	0,25	0,74

La matriz multirasgo-multimétodo se compone de varias submatrices. La matriz superior izquierda contiene las correlaciones entre los tests de la batería A. En la diagonal aparecen los coeficientes de fiabilidad de cada test. Fuera de la diagonal aparecen las correlaciones entre los tests de la batería A. La matriz inferior derecha muestra la misma información referida a la batería B.

La matriz inferior izquierda (sombreada) contiene las correlaciones entre los tests de las dos baterías. En la diagonal están los **coeficientes de validez convergente** (0,65, 0,50 y 0,62), que son las correlaciones entre los dos tests que miden el mismo rasgo. Fuera de la diagonal aparecen los coeficientes de correlación entre distintos rasgos medidos por distintos tests.

Para evaluar los dos tipos de validez mencionados se procede del siguiente modo:

> 1) Los coeficientes de validez convergente deben ser mayores que las correlaciones entre tests que miden diferentes rasgos. En estos datos, existe el problema de que, en la batería A, la correlación entre razonamiento espacial y factor *g* es excesivamente alta, por lo que esta batería no parece discriminar bien entre ambas. Este problema no sucede en la batería B.
>
> 2) El método empleado para medir los rasgos no debe afectar a las relaciones entre ellos. Esto significa que las tres matrices de correlación deben ser similares, exceptuando los elementos de la diagonal. El resultado no es completamente satisfactorio porque la batería B discrimina mejor entre los tres rasgos que la batería A.

4.- VALIDEZ REFERIDA AL CRITERIO

4.1.- CONCEPTO

En el apartado correspondiente al análisis de ítems estudiamos el concepto de índice de validez de un elemento, y ya entonces avanzamos el concepto de **criterio** externo al test, con el que correlacionar el rendimiento en cada ítem.

Sobre todo cuando se pretende utilizar el test para pronosticar determinados criterios de rendimiento (por ejemplo, el rendimiento escolar en un nivel dado, el total de ventas que se van a conseguir, el aprovechamiento de un cursillo o la mejora en un proceso terapéutico) conviene

que el test se relacione muy estrechamente con un criterio externo. Este criterio externo debe ser una medida fiable del rendimiento que se quiere pronosticar con el test: calificaciones escolares, total de ventas producidas en un determinado período, estimaciones de un terapeuta de las mejoras conseguidas por cada persona, etc. A la correlación entre las puntuaciones en el test (X) y en el criterio (Y) se le denomina coeficiente de validez, lo designamos como r_{xy} e indicará el grado en el que el test sirve para pronosticar con precisión el rendimiento en el criterio.

Supongamos, por ejemplo, que la correlación entre un test de conocimientos adquiridos en 1º de BUP y la calificaciones obtenidas en COU es 0,95 en una muestra apropiada. Como la correlación es elevada, cometeríamos errores de pronóstico pequeños, haciendo uso de la oportuna ecuación de regresión, al predecir el rendimiento en COU de un alumno si conocemos su rendimiento en el test. Podríamos estimar con bastante exactitud el rendimiento que manifestará en COU un determinado alumno que se encuentra todavía en 1º de BUP.

El lector puede imaginar que no siempre es útil medir un criterio directamente, debido a razones de coste temporal y económico. Por eso es preciso que los profesionales dispongan de tests con elevada validez relativa al criterio en ámbitos en los que de una u otra forma deben tomar decisiones sobre el nivel de los sujetos en un criterio o sobre su admisión o no a un puesto de trabajo o de estudio determinado.

En muchas ocasiones no resulta sencillo establecer criterios apropiados, fiables y fácilmente mensurables. Los problemas en cualquiera de estas direcciones repercuten disminuyendo el coeficiente de validez y, por tanto, la precisión con que se puede pronosticar un nivel dado en el criterio conociendo la puntuación en el test.

4.2.- INTERPRETACIÓN Y ESTIMACIONES EN EL CRITERIO

El coeficiente de validez es una correlación de Pearson y, por tanto, su interpretación más inmediata se fundamenta en el denominado **coeficiente de determinación**, que es simplemente el cuadrado de la correlación y que indica la proporción de varianza del criterio que podemos pronosticar con el test. Así, un test con un coeficiente de validez de 0.5 indicará que explica un 25 % de la variabilidad o diferencias individuales en el criterio, mientras que el 75 % restante se debe a variables diferentes al test.

Recordando algunos conceptos fundamentales de la regresión lineal simple, el coeficiente de determinación se puede expresar de la siguiente manera:

$$r_{xy}^2 = \frac{S_{y'}^2}{S_y^2} = 1 - \frac{S_{y-y'}^2}{S_y^2}$$

donde S_y^2 es la varianza del criterio

$S_{y'}^2$ es la varianza de los pronósticos

$S_{y-y'}^2$ es la varianza de los errores de pronóstico

La ecuación de regresión de Y sobre X en la escala directa se establece como:

$$Y_i' = (\overline{Y} - r_{xy}\frac{S_y}{S_x}\overline{X}) + r_{xy}\frac{S_y}{S_x}X_i$$

Mediante esta expresión podemos estimar la puntuación directa en el criterio de una determinada persona pero, como es conocido, esa estimación será tanto más precisa cuanto mayor sea la correlación entre test y criterio. Estadísticamente, resulta más apropiada una estimación por intervalos realizada con cierta probabilidad, para lo cual aplicaremos la siguiente expresión:

$$Y_i' \pm Z_{1-\alpha/2}S_{y-y'}$$

donde $Z_{1-\alpha/2}$ es el valor Z, de la normal (0, 1), asociado a la probabilidad establecida y $S_{y-y'}$ es el error típico de estimación.

Ejemplo: A una muestra de 5 alumnos de bachillerato se le aplica un test de habilidades comunicativas (X). A sus respectivos profesores se les pide que hagan una valoración (de 0 a 20 puntos) de la capacidad de relación interpersonal de sus alumnos. Estas valoraciones hacen la función de criterio (Y). Los resultados en el test y en el criterio fueron los siguientes:

Alumno	X	Y
1	7	6
2	13	10
3	10	9
4	9	8
5	11	12
Media	10	9
Varianza	4	4

El coeficiente de validez del test es r_{xy} = 0.8, lo que significa que el test de habilidades comunicativas explica un 64 % de las diferencias en las valoraciones de los profesores sobre la capacidad de relación interpersonal de sus alumnos.

Si queremos pronosticar puntualmente la puntuación en el criterio del alumno nº 5, aplicando la oportuna ecuación de regresión obtenemos:

$Y_5' = 9.8$

Para realizar la estimación por intervalo para este mismo alumno, con probabilidad 0.95, fijamos el valor $Z_{1-\alpha/2} = 1.96$ y calculamos el error típico de estimación:

$$S_{y-y'} = S_y\sqrt{1-r_{xy}^2} = 1.2$$

y el intervalo será:

$$9.8 \pm (1.96)\,(1.2) \begin{cases} \nearrow 12.152 \\ \searrow 7.448 \end{cases}$$

Diremos entonces, con probabilidad 0.95, que la puntuación del alumno 5 en el criterio se encontrará entre 12.152 y 7.448.

Cuando, tanto en contextos aplicados como investigadores, se desea predecir de la forma más precisa posible las puntuaciones en un determinado criterio, es común utilizar más de un predictor. En este caso debe aplicarse la técnica estadística de Regresión Múltiple, que proporciona los pesos (coeficientes de regresión parcial) de cada predictor según la importancia que tengan para la predicción.

4.3.- FACTORES QUE AFECTAN AL COEFICIENTE DE VALIDEZ

Centrándonos en la validez relativa al criterio, el coeficiente de validez es una correlación entre una variable X (test) y otra Y (criterio). La cuantía de la correlación viene condicionada por varios factores, como son:

- La fiabilidad del test.
- La fiabilidad del criterio.
- La auténtica relación entre test y criterio.
- La variabilidad de la muestra en el test y en el criterio.

Respecto a los dos primeros factores, aunque no tratamos en toda su extensión el desarrollo formal de las relaciones, podemos decir que el coeficiente de validez tiende a incrementarse a medida que test y criterio son variables medidas con exactitud. Problemas de fiabilidad en uno u otro se reflejan mediante una disminución del coeficiente de validez. De hecho, se puede comprobar que el límite máximo al que puede llegar r_{xy} es $\sqrt{r_{xx}r_{yy}}$. Es decir,

$$r_{xy} \leq \sqrt{r_{xx}r_{yy}}$$

siendo r_{xx} el coeficiente de fiabilidad del test y r_{yy} el coeficiente de fiabilidad del criterio.

Demostración:

Una de las expresiones de la correlación de Pearson es:

$$r_{xy} = \frac{\sum (x-\bar{x})(y-\bar{y})}{NS_xS_y}$$

Realizando las sustituciones oportunas que permiten los supuestos de la Teoría Clásica:

$$r_{xy} = \frac{\sum (V_x + E_x - \bar{V}_x)(V_y + E_y - \bar{V}_y)}{NS_xS_y}$$

Si realizamos los productos término a término en el numerador, divididos entre N resultan covarianzas, y finalmente, el coeficiente de validez quedaría como:

$$r_{xy} = \frac{Cov(V_x, V_y)}{S_xS_y}$$

Ya que el resto de covarianzas del numerador, haciendo uso de los supuestos del modelo clásico, son cero.

Otra manera de expresar la ecuación anterior es:

$$r_{xy} = \frac{r_{V_xV_y}S_{V_x}S_{V_y}}{S_xS_y} = r_{V_xV_y}\sqrt{r_{xx}r_{yy}}$$

Dado que la correlación entre puntuaciones verdaderas entre el test y puntuaciones verdaderas en el criterio es igual o inferior a 1, queda demostrada la desigualdad.

Imaginemos, por ejemplo, que un test de inteligencia general manifiesta un $r_{xx} = 0.85$, mientras que una prueba de cultura general, considerada como criterio, manifiesta un $r_{yy} = 0.73$. Según

estos datos, el coeficiente de validez de este test respecto a este criterio no puede superar el valor de 0.79, que es la raíz cuadrada del producto entre los dos coeficientes de fiabilidad.

De lo anterior se deduce que el coeficiente de validez de un test es menor o igual que la raíz cuadrada del coeficiente de fiabilidad del test; también es menor o igual que la raíz cuadrada de la fiabilidad del criterio:

$$r_{xy} \leq \sqrt{r_{xx} r_{yy}} \leq \sqrt{r_{xx}}$$

$$r_{xy} \leq \sqrt{r_{xx} r_{yy}} \leq \sqrt{r_{yy}}$$

dado que el valor máximo de un coeficiente de fiabilidad es uno.

Por otra parte, y atendiendo ahora a las relaciones entre longitud del test y su fiabilidad, es lógico que si la fiabilidad influye directamente en el coeficiente de validez, la longitud del test (y en su caso, del criterio) influya también en r_{xy} aunque de modo indirecto. La fórmula que permite estimar el coeficiente de validez de un test alargado n veces (compuesto por n formas paralelas) es:

$$R_{xy} = \frac{r_{xy}}{\sqrt{\frac{1-r_{xx}}{n} + r_{xx}}}$$

donde R_{xy} es el coeficiente de validez del test alargado.
r_{xy} es el coeficiente de validez del test original.
r_{xx} es el coeficiente de fiabilidad del test original.
n es el nº de veces que se alarga el test original.

Demostración:

Sean r_{xy} r_{xx} y r_{yy}, respectivamente, los coeficientes de validez, de fiabilidad del test y de fiabilidad del criterio. Supongamos que alargamos con formas paralelas la longitud del test, con lo cual aumentarán su coeficiente de fiabilidad (R_{xx}) y su coeficiente de validez (R_{xy}), mientras que el criterio (que no se modifica) permanece con el mismo coeficiente de fiabilidad.

Según las relaciones vistas anteriormente, podemos establecer las siguientes igualdades, para el coeficiente de validez del test inicial y del test alargado:

$$r_{xy} = r_{V_x V_y} \sqrt{r_{xx} r_{yy}} \quad \text{y} \quad R_{xy} = r_{V_x V_y} \sqrt{R_{xx} r_{yy}}$$

Dividiendo miembro a miembro y despejando el coeficiente de validez del test alargado, tendríamos que:

$$R_{xy} = \frac{r_{xy}}{\sqrt{\frac{r_{xx}}{R_{xx}}}} = \frac{r_{xy}}{\sqrt{\frac{r_{xx}}{nr_{xx} \Big/ \left(1+(n-1)r_{xx}\right)}}} = \frac{r_{xy}}{\sqrt{\frac{1-r_{xx}}{n} + r_{xx}}}$$

Ejemplo: Supongamos que una "Escala de actitud hacia grupos ecologistas" de 30 items manifiesta en un grupo normativo un coeficiente de fiabilidad de 0,42 y un coeficiente de validez de 0,51. Si se duplicase le longitud de la escala, es decir si se le añadiera una forma paralela de 30 items, el coeficiente de validez pasaría a valer:

$$R_{xy} = \frac{0,51}{\sqrt{\frac{1-0,42}{2} + 0,42}} = 0,60$$

Si de la fórmula anterior despejamos *n*, podemos estimar el número de veces que deberemos multiplicar la longitud del test para alcanzar un coeficiente de validez R_{xy} deseado:

$$n = \frac{1-r_{xx}}{\frac{r_{xy}^2}{R_{xy}^2} - r_{xx}}$$

En caso de que el valor de *n* sea negativo, significa que el valor deseado no es alcanzable incrementando la longitud del test.

En el caso hipotético de un test infinitamente largo o, lo que es lo mismo, de un test con máxima precisión, en la siguiente fórmula, R_{xx} valdría 1, y R_{xy} se podría interpretar como el máximo coeficiente de validez obtenible como resultado de mejorar la fiabilidad del test todo lo posible.

$$R_{xy} = \frac{r_{xy}}{\sqrt{\frac{r_{xx}}{R_{xx}}}} = \frac{r_{xy}}{\sqrt{\frac{r_{xx}}{1}}} = \frac{r_{xy}}{\sqrt{r_{xx}}}$$

Ejemplo: Un determinado test de 10 items manifiesta en un grupo normativo un coeficiente de fiabilidad de 0.4 y un coeficiente de validez de 0.35. Nos cuestionamos cuántos items paralelos necesitaría el test para conseguir:

a) Un coeficiente de validez de 0.5

b) Un coeficiente de validez de 0.9

a)

$$n = \frac{1-0.4}{\frac{0.35^2}{0.5^2}-0.4} = 6.7$$

b)

$$n = \frac{1-0.4}{\frac{0.35^2}{0.9^2}-0.4} = -2.4$$

Podemos comprobar a partir de estos cálculos que el coeficiente de validez de 0.5 lo conseguiremos con un test de, aproximadamente, 70 items; con lo cuál habría que diseñar 6 formas adicionales paralelas al test original.

El coeficiente de validez de 0.9 es imposible de conseguir, por mucho que incrementemos la longitud del test inicial con formas paralelas, de ahí que en "b", hallamos obtenido un valor de *n* negativo. El máximo coeficiente de validez obtenible mejorando la fiabilidad (o alargando el test) es $R_{xy} = r_{xy} / \sqrt{r_{xx}} = 0.35 / \sqrt{0.4} = 0.55$, que es menor de 0.9.

Hemos indicado también que r_{xy} depende de la variabilidad del grupo normativo. De forma parecida a lo que acontecía con la varianza del grupo en el test y su coeficiente de fiabilidad, el coeficiente de validez de un test respecto a un criterio es tanto más elevado cuanto mayor es la varianza de grupo normativo en ambos. Significa esto que, por ejemplo, un test de aptitud para la venta tendrá un coeficiente de validez mayor en una muestra de la población general (donde habrá heterogeneidad respecto a la aptitud por ser vendedor) que en una muestra de vendedores experimentados (seguramente obtendrían todos puntuaciones elevadas, y por tanto sería un grupo más homogéneo). En la medida que el poder predictivo de un test respecto a un criterio depende de su r_{xy}, habrá que considerar la variabilidad del grupo donde se ha obtenido.

5.- ALGUNOS EJEMPLOS EMPÍRICOS DEL PROCESO SEGUIDO PARA LA VALIDACIÓN DE TESTS

En las siguientes páginas mostramos algunos trabajos desarrollados para la validación de varios tests psicológicos, de contenido y objetivos bien diversos. Hemos intentado incluir ejemplos que sigan estrategias de investigación diferentes para obtener información sobre el constructo que se mide o sobre el tipo de inferencias que se pueden hacer a partir de las puntuaciones obtenidas en los tests.

5.1.- Barraca, J., López-Yarto, L. & Olea, J. (2000). Psychometric properties of a new Family Life Satisfaction Scale. *European Journal of Psychological Assessment*, 16, 2, 98-106.

Los autores elaboraron una nueva escala o cuestionario para evaluar la satisfacción familiar. Argumentan que se ha hecho poco esfuerzo por definir este constructo desde un marco teórico concreto, lo que ha dado lugar a instrumentos de evaluación de la satisfacción familiar fundamentados en una pobre definición del constructo. Los trabajos sobre instrumentos previos de evaluación han estudiado su relación con otras variables (por ejemplo, con la satisfacción hacia la calidad de vida, con el constructo "locus of control" o con el nivel de religiosidad) que al menos puede decirse que son cuestionables. Critican también que los instrumentos hasta entonces disponibles no incluyen suficientemente los componentes afectivos del constructo. Además, algunos de los cuestionarios previos para evaluar la satisfacción familiar resultan poco amigables de responder: uno de ellos, por ejemplo, consiste en preguntar dos veces sobre los mismos temas, una vez sobre la situación real de su familia y otra sobre lo que sería su familia ideal.

Todo ello les lleva a la opción de construir una nueva escala de satisfacción familiar, para lo cual siguieron el siguiente procedimiento:

Definición del constructo: Se entiende la satisfacción familiar como el conjunto de sentimientos que cada persona experimenta en su propia familia, y que son el resultado de sus continuas interacciones con los demás, así como de las consecuencias positivas o negativas derivadas.

Instrumento inicial de evaluación: Decidieron evaluar estas connotaciones afectivas mediante una escala de adjetivos bipolares, también denominado diferencial semántico, que tenía el siguiente formato:

Cuando estoy en casa con mi familia, normalmente me siento:

Feliz	_____	_____	_____	_____	_____	_____	_____	*Infeliz*
Solo	_____	_____	_____	_____	_____	_____	_____	*Acompañado*

Cada ítem se puntuó desde uno hasta 7, dado que había ese número de categorías ordenadas de respuesta. Inicialmente elaboraron 177 adjetivos bipolares y eliminaron 66 por resultar redundantes. Tres especialistas en terapia de familia dejaron la lista en 52, aquellos que de forma unánime fueron considerados relevantes para evaluar el constructo.

Análisis y selección de ítems: Se aplicó la escala inicial a una muestra de 274 personas. Mediante el programa SPSS se obtuvieron varios indicadores psicométricos para cada uno de los 52 ítems: a) correlación ítem-total, b) varianza, c) saturaciones factoriales (rotación varimax), y d) coeficiente α de la escala cuando se elimina el ítem. Se retuvieron finalmente los 27 ítems que cumplieron simultáneamente los siguientes requerimientos: a) correlación ítem-total mayor de 0.45, b) varianza por encima de 1, c) saturaciones en el primer factor rotado por encima de 0.30, y d) coeficiente α de la escala (al eliminar el ítem) igual o superior al de la escala completa (α=0.9808).

Estudio de la fiabilidad: Se obtuvo un coeficiente α igual a 0.976. El coeficiente de fiabilidad test-retest, obtenido tras un período de 4 semanas, resultó ser 0.758; aún no siendo óptimo este coeficiente, es bastante usual que la estabilidad temporal no sea mucho más alta cuando se emplea un diferencial semántico como instrumento de evaluación.

Validez de constructo factorial: Se realizó un nuevo análisis factorial sobre la matriz de correlaciones entre los 27 ítems. El primer factor explicó el 62.3 % de la varianza total, lo que se consideró suficiente prueba de unidimensionalidad. Todos los ítems obtuvieron saturaciones por encima de 0.68 en el primer factor sin rotar.

Validez convergente: Se aplicaron a la misma muestra dos de los instrumentos previos de evaluación: el cuestionario *Family Satisfaction* (Olson y Wilson, 1982) y la *Family Satisfaction Scale* (Carver y Jones, 1992). El nuevo cuestionario correlacionó 0.646 con las puntuaciones en el primero y 0.787 con las correspondientes en el segundo.

Datos adicionales sobre la validez de constructo: Se aplicó el nuevo cuestionario a una muestra de 16 personas (con la misma edad media de la muestra general) que asistían a una terapia de familia. La media de esta muestra clínica en el cuestionario fue de 97.56, mientras que la media de la muestra general fue 121.56. El contraste estadístico entre ambas medias (prueba U de Mann Whitney) resultó significativo con un nivel de confianza del 95 %, con lo que se concluyó que la nueva escala era capaz de diferenciar el grado de satisfacción familiar de ambas muestras.

5.2.- Ehlers, S., Gillberg, Ch. & Wing, L. (1999). A screening questionnaire for Asperger Syndrome and other High-Functioning Autism Spectrum disorders in school age children. *Journal of Autism and Developmental Disorders*, 29, 2, 129-141.

En el presente artículo se describe un estudio realizado para comprobar las propiedades psicométricas de un nuevo instrumento, el *Autism Spectrum Screening Questionnaire (ASSQ)*, diseñado para detectar (no tanto evaluar con precisión) a chicos y chicas que tienen severos desordenes autistas pero con alto funcionamiento cognitivo, en concreto el denominado como "síndrome de Asperger". Este alto funcionamiento cognitivo complica mucho la detección de esta patología.

Descripción del síndrome: No existiendo un acuerdo universal sobre los síntomas del trastorno de Asperger, parece que se trata de chicos sin demasiados retrasos en el lenguaje ni

en su desarrollo cognitivo, pero que tienen síntomas claramente autistas en lo que se refiere a problemas de interacción social y de conductas estereotipadas.

Elaboración del cuestionario: Varios especialistas clínicos ingleses y suecos elaboraron un listado de síntomas característicos del síndrome en chicos de entre 7 y 16 años. Ellos mismos elaboraron 27 ítems que recogieran esos síntomas y que fueran inteligibles para personas no expertas (padres y profesores), ya que no intentaban tanto diagnosticar con precisión el síndrome como que informantes no expertos (padres o profesores) identificaran a los chicos que necesitaban un diagnóstico diferencial en profundidad. La sintomatología que pretendían incluir era: interacción social, problemas de comunicación, conducta repetitiva y esterotipias motoras. El formato de ítems y respuesta que establecieron fue:

Este chico destaca como diferente de otros chicos de su edad en los siguientes aspectos:

- Carece de sentido común	*No*	*Algo*	*Sí*
- Carece de empatía	*No*	*Algo*	*Sí*
- Tiene movimientos involuntarios en la cara o el cuerpo	*No*	*Algo*	*Sí*

Cada respuesta era cuantificada como 0, 1 ó 2 puntos, con lo que el rango teórico de puntuaciones podía oscilar entre 0 y 54.

Muestras seleccionadas: En el estudio se describe la selección de dos muestras de chicos diagnosticados previamente con determinados desórdenes conductuales por diversos psicólogos y psiquiatras. La muestra principal estaba formada por 3 tipos de patologías: 21 casos de desórdenes de espectro autista (en el que se incluye el síndrome de Asperger), 58 casos con déficit atencional, hiperactividad y conducta disruptiva, y 31 con problemas de aprendizaje (retraso en lectura y escritura). La muestra de validación estaba formada por 34 chicos y chicas diagnosticados previamente en contextos clínicos como síndromes de Asperger.

Fiabilidad. La fiabilidad test-retest, con dos semanas de diferencia entre las dos aplicaciones, fue 0.96 cuando los evaluadores eran los padres y 0.94 cuando eran los profesores.

La correlación entre las evaluaciones de los padres y de los profesores (fiabilidad interjueces) se obtuvo en los tres grupos de la muestra principal. Considerando la evaluación de la muestra completa, esta correlación fue 0.66, mientras que resultó 0.77 para los chicos con espectro autista, 0.27 para los chicos con déficit atencional y 0.19 para los chicos con trastornos de aprendizaje.

Validez convergente. Los padres y profesores respondieron también a dos escalas generales de evaluación de psicopatologías en niños, las escalas de Rutter y las de Conners, obteniendo correlaciones de 0.75 y 0,58, respectivamente, en la muestra de padres, así como valores de 0.77 y 0.70 en la muestra de profesores.

Validez referida al criterio. En este caso, uno de los objetivos fundamentales del trabajo consistía en estudiar el grado en que las puntuaciones totales en el cuestionario ASSQ servía para diferenciar a los diversos grupos diagnósticos que formaban la muestra principal. Se realizaron los correspondientes ANOVAS, donde la variable independiente era el grupo

diagnóstico y la variable dependiente las puntuaciones en un cuestionario concreto (ASSQ, Rutter o Conners). Algunos resultados interesantes fueron: a) los tres grupos de la muestra principal obtuvieron puntuaciones medias significativamente distintas en el cuestionario ASSQ, tanto cuando los evaluadores eran padres como cuando eran profesores; b) los chicos con diagnóstico de espectro autista obtuvieron siempre las medias más elevadas; c) las puntuaciones en las otras dos escalas no consiguieron diferencias significativas entre los chicos de espectro autista y los hiperactivos; d) las medias de las puntuaciones (asignadas por ambos tipos de evaluadores) en el cuestionario ASSQ, fueron estadísticamente similares en la muestra de validación (síndrome Asperger) que en la submuestra de espectro autista de la muestra principal.

Establecimiento de puntos de corte. Antes de comprender lo que realmente se hizo en este trabajo, conviene describir el procedimiento general y su sentido. En contextos de diagnóstico clínico interesa muchas veces estudiar el grado en que las puntuaciones en un cuestionario sirven para clasificar de forma fiable, es decir, si sirve el cuestionario para clasificar correctamente a una persona dentro o fuera del grupo clínico objeto de estudio. Así, podríamos establecer una determinada puntuación como punto de corte, de tal manera que si la puntuación de una persona supera dicho punto de corte la clasificaríamos en el grupo con trastornos, mientras que si se encuentra por debajo de dicho punto de corte concluiríamos que no tiene dicho trastorno. Para establecer un determinado punto de corte, debemos entender en primer lugar dos conceptos esenciales:

- La *sensibilidad*, también denominada probabilidad de acierto o de verdaderos positivos, que es la proporción de personas realmente diagnosticadas con desórdenes que las clasificamos como tales mediante el cuestionario.
- La *especificidad* o proporción de personas sin trastorno que los clasificamos como tales a partir de sus puntuaciones en el cuestionario. La proporción complementaria a la especificidad es la *probabilidad de falsos positivos* (también denominadas como falsas alarmas), que es la proporción de personas que realmente no tienen el trastorno y que decimos a partir del cuestionario que sí lo tienen.

Las dos proporciones anteriores variarán según la puntuación total en el cuestionario que establezcamos como punto de corte (en el caso del ASSQ podríamos establecer en teoría hasta 55 puntos de corte diferentes). Por ejemplo, si en el ASSQ pusiéramos como punto de corte la puntuación 54, que es la máxima posible, obviamente la sensibilidad sería 0 (todos los chicos con síndrome Asperger quedarían clasificados como no Asperger) y la especificidad 1 (todos los chicos sin síndrome Asperger quedarían clasificados como tales); si el punto de corte lo pusiéramos en la puntuación 0, la sensibilidad sería 1 pero la especificidad 0. Por tanto, para decidir sobre el punto de corte más apropiado tendríamos que intentar maximizar ambas proporciones simultáneamente, lo cual depende del grado de validez predictiva de las puntuaciones del cuestionario para diferenciar los dos diagnósticos posibles. En la práctica, para cada puntuación posible como punto de corte, suele representarse en un cuadrado unitario la proporción complementaria a la especificidad o proporción de falsos positivos (en el eje de abcisas) y la sensibilidad (en el eje de ordenadas). Esta representación se conoce como curva ROC, y muchas veces interesa establecer como punto de corte aquella puntuación del cuestionario que queda representada más cerca de la esquina superior izquierda del cuadrado unitario. Esa puntuación será la que maximiza simultáneamente la especificidad y la sensibilidad.

En el presente estudio, se obtuvieron, por ejemplo, las siguientes proporciones de sensibilidad y de falsos positivos, cuando los informantes eran los padres y el trastorno era el referido a desordenes de espectro autista (subgrupo de la muestra principal):

Punto de corte	Sensibilidad	Falsos positivos	Cociente entre ellos
7	.95	.44	2.2
13	.91	.23	3.8
15	.76	.19	3.9
16	.71	.16	4.5
17	.67	.13	5.3
19	.62	.10	5.5
20	.48	.08	6.1
22	.43	.03	12.6

Los autores indican que si se valora mucho la sensibilidad, es decir, intentar no cometer errores con los chicos realmente diagnosticados como autistas, sería aconsejable establecer como punto de corte la puntuación X=13, a partir de la cual se detecta al 91 % de los chicos con autismo de la muestra principal. El coste de esta clasificación es que clasificaríamos como autistas a un 23 % de los chicos de la muestra principal que son hiperactivos o tienen problemas de aprendizaje. Tal coste no sería muy elevado si el cuestionario representa sólo una primera detección, y es posible posteriormente profundizar en el diagnóstico diferencial mediante procedimientos alternativos.

Sin embargo, si el objetivo realmente fuera distinguir entre los chicos autistas y los que tienen otros trastornos, los autores optan por establecer un punto de corte en la puntuación X=19, lo cual minimiza la proporción de verdaderos positivos (0.62) pero también la de falsos negativos (0.10). Teniendo en cuenta el tamaño muestral, esta decisión equivale a emitir un 82 % de decisiones correctas.

5.3.- Olea, J., Abad, F.J. y Ponsoda, V. (2002). Elaboración de un banco de ítems, predicción de la dificultad y diseño de anclaje. *Metodología de las Ciencias del Comportamiento*, Vol. Especial, 427-430.

Olea, J., Abad, F.J., Ponsoda, V. y Ximénez, M.C. (2004). Un test adaptativo informatizado para evaluar el conocimiento del inglés escrito: Diseño y comprobaciones psicométricas. *Psicothema* 16, 519-525.

En ambos trabajos se recogen los estudios realizados para poner en funcionamiento un Test Adaptativo Informatizado (TAI) de conocimientos del idioma inglés en su versión escrita. Este tipo de tests requieren un amplio banco de ítems, su estudio psicométrico desde la Teoría de la Respuesta al Ítem, así como un conjunto de programas informáticos para la presentación de los mejores ítems a cada persona y para la estimación de su nivel (normalmente entre un rango de valores que oscila entre –4 y +4). Nos centraremos fundamentalmente en el proceso de construcción del banco de ítems, que conforma el contenido fundamental del TAI, y en el análisis de sus propiedades psicométricas.

Criterios generales para la elaboración del banco de ítems. Varias especialistas en Filología Inglesa, junto a varios profesionales de la Psicometría, elaboraron el banco de ítems. Los psicómetras indicaron a las filólogas algunos criterios a considerar en la elaboración del banco de ítems: a) debía tener aproximadamente 600 ítems, b) su dificultad previsible debía ser heterogénea, ya que el TAI pretende evaluar cualquier nivel de dominio del inglés escrito, d) los ítems debían ser de opción múltiple, siendo el enunciado una frase donde faltarían ciertas palabras, y 4 opciones de respuesta de las que sólo una es correcta, e) las 3 opciones incorrectas de un ítem debían cumplir los requisitos de redacción que son aconsejables (ver tema 1 de estos materiales), f) las filólogas deberían partir de un modelo teórico (ellas dirían cual) explicativo de lo que representa el conocimiento del inglés escrito.

Validez de contenido del banco de ítems. Las especialistas en Filología establecieron un modelo de dominio del inglés escrito funcional-cognitivo, en el que además de la competencia gramatical (aspectos fundamentalmente sintácticos) se incluyeran contenidos para evaluar la competencia en el discurso (componentes pragmáticos y léxicos), de tipo más instrumental para contextos comunicativos concretos. Finalmente establecieron 7 categorías gramaticales generales, denominadas como: aspectos formales, componentes morfológicos, sintácticos, morfosintácticos, pragmáticos, léxicos y una categoría mixta. Estas categorías se dividían a su vez en otras subcategorías: por ejemplo, en los componentes morfológicos (222 ítems en total) se especificaron 17 subcategorías diferentes. De esta forma se redactaron un total de 635 ítems, cada uno perteneciente a una categoría y subcategoría específicas. Un ejemplo de un ítem de morfología (subcategoría de tiempos verbales) es:

*We * when he gets here.*
a) wouldn´t b) have left c) couldn´t leave d) will have left

Cinco personas nativas, profesores de inglés en diversos centros educativos, revisaron el banco, detectando errores de diverso tipo en la redacción y marcando lo que consideraban como respuestas correctas. Varias reuniones entre una de las filólogas y uno de los nativos sirvieron para corregir los errores y no tener duda sobre la opción correcta de algunos ítems.

Elaboración de subtests equivalentes. Para estudiar las propiedades psicométricas del banco de ítems es necesario aplicarlo a muestras de personas con nivel heterogéneo de inglés. Sin embargo, es prácticamente imposible aplicar 635 ítems a cada persona, con lo cual es necesario establecer lo que se denomina un "diseño de anclaje", que consiste en construir subtests, de tal forma que sean lo más equivalentes posible en dificultad y en contenidos incorporados. Así, se elaboraron 15 subtests diferentes, cada uno de los cuales tenía las siguientes características:

a) Estaba formado por 61 ítems, 41 propios de ese subtest y 20 comunes a todos los subtests.
b) Tenía una dificultad heterogénea. Los 5 nativos habían valorado subjetivamente la dificultad de los ítems, y en cada subtest se incluían ítems de amplia gama de dificultad previsible.
c) La cantidad de ítems de una categoría se decidía según el peso de esa categoría en el banco completo; por ejemplo, para la categoría "morfología", que tenía el 35 % de los ítems del banco (222 de los 635 ítems totales), se eligieron 20 ítems para cada subtest (aproximadamente el 35 % de 61).

Estudio piloto de uno de los subtests. El primero de los subtests se aplicó a una muestra de 435 personas: estudiantes de ESO y Bachillerato, estudiantes y profesores de Psicología, y estudiantes de Filología Inglesa. Se les pidió alguna información adicional, como su autoevaluación del dominio del inglés y el procedimiento seguido para el aprendizaje del idioma (colegio, familia, escuela oficial de idiomas,etc.). Algunos resultados de este primer estudio psicométrico fueron:

a) Se eliminaron 9 ítems por ser demasiado fáciles o correlacionar de forma escasa con el total del subtest.
b) De los 52 ítems retenidos, la media de las correlaciones ítem-total fue 0.556. El coeficiente α de Cronbach resultó ser de 0.91
c) El análisis factorial sobre la matriz de correlaciones tetracóricas entre los 52 ítems dio lugar a un factor con varianza explicada de 15.78 (30.35 % de la varianza total), lo que se consideró como prueba suficiente de unidimensionalidad. Esto es un requisito para la aplicación del modelo de TRI seleccionado por los investigadores.
d) Se realizó un análisis de regresión múltiple, donde las variables independientes fueron la información adicional recogida y la variable dependiente el nivel de conocimientos estimado desde la TRI. El coeficiente de correlación múltiple entre las variables adicionales (autoevaluación y formación en el idioma) y las puntuaciones estimadas en el subtest resultó ser 0.747.

Estudio psicométrico simultáneo de 5 subtests. Se aplicaron 5 de los subtests restantes a una muestra de 3224 estudiantes de primer curso de la Pontificia Universidad Católica de Chile. Cada grupo de algo más de 600 estudiantes respondió a uno de los subtests y a una serie de preguntas adicionales de un cuestionario, donde se recabó información sobre el tipo de colegio donde estudiaron la enseñanza media (bilingüe-inglés u otros), su nivel autopercibido en el idioma (en escritura, lectura y conversación) y sobre su formación complementaria en el idioma (en casa, en estancias prolongadas en países anglófonos, en escuelas oficiales de idiomas, etc). Además de obtener de nuevo información sobre la consistencia interna y unidimensionalidad de los subtests (resultados satisfactorios para ambos objetivos) se estudió la validez predictiva de las puntuaciones. Más concretamente, se realizaron dos estudios:

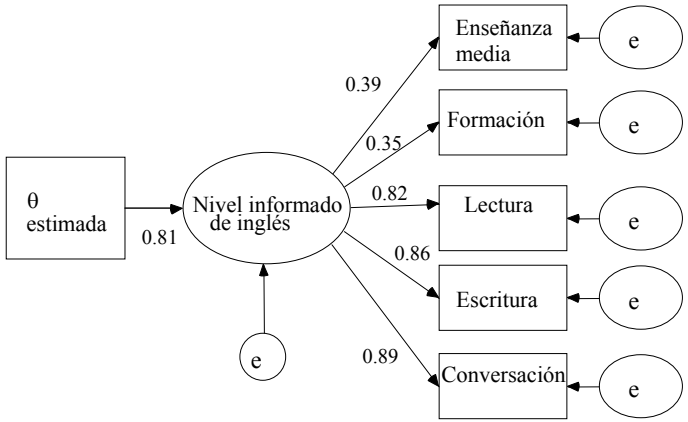
a) Las primeras pruebas de validez se realizaron a partir de los datos obtenidos en el cuestionario. Se realizaron 5 ANOVAs, uno por cada variable independiente incluida en el cuestionario, siendo en todos ellos la variable dependiente el nivel de rasgo estimado (en una escala de -4 a 4) para cada estudiante a partir de sus respuestas al subtest correspondiente (los cinco valores F resultaron significativos, $p < 0.001$): a) con la variable independiente *tipo de colegio*, los niveles de rasgo medios (θ) fueron 0.50 (colegio bilingüe-inglés) y -0.24 (otros colegios). El tamaño del efecto (η^2) fue 0.10. b) con la variable independiente *formación*, los niveles de rasgo medio fueron -0.16 (sólo colegio), 0.24 (colegio+academia), 0.57 (colegio+familia) y 1.18 (colegio+extranjero). El tamaño del efecto fue 0.09. c) con la variable independiente *autoevaluación de la lectura*, los niveles de rasgo medio fueron -1.16 (nada), -0.86 (sencillo), -0.13 (con esfuerzo), 0.94 (bien) y 1.64 (bilingüe). El tamaño del efecto fue 0.46. d) con la variable independiente *autoevaluación de la escritura*, los niveles de rasgo medio fueron -1.30 (nada), -0.64 (sencillo), 0.03 (con esfuerzo), 0.90 (bien) y 1.77 (bilingüe). El tamaño del efecto fue 0.49. e) con la variable independiente *autoevaluación de la conversación*, los niveles de rasgo medio fueron -1.23 (nada), -0.66 (sencillo), 0.25 (con esfuerzo), 1.01 (bien) y 1.76 (bilingüe). El tamaño del efecto fue 0.53. En los cinco análisis se

observa que los niveles de rasgo medios se incrementan a medida que lo hacen los niveles de cada una de las variables independientes. Todas las comparaciones múltiples post hoc (estadístico DHS de Tukey) resultaron significativas ($p<0.05$). En los valores de los tamaños del efecto (η^2) puede observarse un mayor poder predictivo de las autoevaluaciones del nivel de inglés que de las variables relacionadas con la formación en el idioma.

b) Adicionalmente se puso a prueba mediante el programa AMOS (versión 4.01) un modelo estructural para obtener la capacidad predictiva de las estimaciones de los niveles de conocimiento con relación a una variable latente de nivel informado de inglés, donde tuvieran saturaciones positivas las 5 variables evaluadas en el cuestionario. Este tipo de análisis estadístico, denominado como "ecuaciones estructurales" o también como "modelos confirmatorios" (véase Ruiz[3], 2000), sirve para estudiar el grado de ajuste entre un modelo teórico (donde se plantean ciertas relaciones entre variables empíricas y teóricas) y los datos reales. En nuestro caso, el modelo teórico consiste en plantear una variable latente o factor (nivel informado de inglés) en la que obtuvieran saturaciones positivas las 5 variables incluidas en el cuestionario; además, planteamos una relación positiva entre esta variable latente y las puntuaciones estimadas a partir del rendimiento manifestado en los subtests de inglés escrito.

Algunas medidas de ajuste del modelo fueron: χ^2/gl = 4.599, AGFI = 0.992, RMSEA = 0.037, que son indicadores de un buen ajuste del modelo teórico a los datos empíricos. Las estimaciones de las saturaciones se recogen en la siguiente figura. Puede comprobarse que la correlación entre las estimaciones de nivel de inglés y el factor latente de nivel informado de inglés es 0.81.

[3] Ruiz, M.A. (2000). Introducción a los modelos de ecuaciones estructurales. Madrid: UNED Ediciones.



EJERCICIOS

1. Señale el objetivo que se pretende conseguir con cada una de las siguientes actuaciones en la construcción de un cuestionario.

a) Correlacionar las puntuaciones totales en el cuestionario con un criterio externo al test.
b) Preguntar a varios jueces expertos sobre la representatividad de los contenidos de un test.
c) Aplicar un análisis factorial a las puntuaciones obtenidas en el test y en varios tests relacionados con el constructo de interés.

2. Señale qué variables pueden afectar al coeficiente de validez de un test (r_{xy}).

3. Sabemos que aumentando la longitud de un test, podemos aumentar también su fiabilidad, y que la fiabilidad del test es un factor que permite incrementar la validez del test. Queremos obtener un coeficiente de validez de 0,8 (R_{xy}) y sabemos que la fiabilidad del test es 0,8 (r_{xx}) y la del criterio es 0,6 (r_{yy}). ¿Lograremos nuestro objetivo aumentando la fiabilidad del test?

4. Un psicólogo social diseña un test con 5 ítems y obtiene los coeficientes de fiabilidad, r_{xx}= 0,4, y validez, r_{xy}= 0,36. En vista de estos valores tan bajos, decide rechazar el test. Valore esta actuación del psicólogo.

5. A continuación se detallan las puntuaciones que 10 personas obtuvieron en un test de rendimiento escolar (X) y las calificaciones medias del curso (Y), que se consideran como un criterio de aprovechamiento académico.

Sujetos	1	2	3	4	5	6	7	8	9	10
Test	18	15	12	11	8	4	5	6	9	3
Calificación	9	8	7	6	4	2	4	4	5	2

a) Suponiendo que las 10 personas constituyen un grupo normativo apropiado, obtenga el coeficiente de validez del test.
b) Obtenga el error típico de estimación del test.

6. En un test de 10 ítems el coeficiente de fiabilidad es 0.25 y el de validez es 0.10.
a) Correlacionando las puntuaciones del test con otro criterio, ¿podríamos obtener un coeficiente de validez menor de 0.10? Razone su respuesta.

b) Correlacionando las puntuaciones del test con otro criterio distinto, ¿podríamos obtener un coeficiente de validez mayor de 0.60? Razone su respuesta.

7. Un pequeño test de aptitudes intelectuales consta de dos ítems de aptitud verbal (el 1 y el 4) y de dos ítems de aptitud numérica (el 2 y el 3). Después de aplicarse a un grupo normativo, la matriz de correlaciones se sometió a un análisis factorial, cuya matriz F rotada se presenta en la tabla que aparece a continuación.

a) ¿Considera que el estudio factorial aporta datos a la validez del test?
b) Calcule el porcentaje de la varianza total explicado por el Factor I.

Ítem	Factor I	Factor II
1	0,247	0,883
2	0,906	0,083
3	0,937	0,024
4	-0,108	0,925

8. Estamos intentando elaborar una escala que mida la calidad de ciertos productos. Cada ítem consiste en un adjetivo y la persona ha de evaluar de "1" (totalmente en desacuerdo) a "5" (totalmente de acuerdo) en qué medida el adjetivo se aplica al producto. Tras un análisis factorial, la matriz rotada resultante ha sido:

	Factor I	Factor II	Factor III
Barato	-0,1	0,8	0,2
Agradable	0,2	0,2	0,7
Útil	0,9	-0,1	-0,2
Cómodo	0,6	-0,2	-0,1
Precio justo	0,1	0,9	0,2
Bonito	-0,2	-0,1	0,6
Necesario	0,7	0,2	0,0
Atractivo	-0,2	-0,1	0,5
Práctico	0,8	0,1	0,2
Manejable	0,8	0,2	-0,1

¿Qué aspectos de la calidad mide la escala?

9. En la selección de aspirantes a un curso de formación, los sujetos han sido examinados con un cuestionario que obtuvo una media de 5 y una desviación típica de 2. Una vez terminado el curso, los mismos sujetos fueron valorados por sus formadores según una escala de 0 a 20, con media 10 y desviación típica 3. La correlación entre los resultados en el cuestionario y las valoraciones de los formadores fue de 0.35.

a) Realice una estimación puntual de la valoración que recibiría una persona que obtuvo 4 puntos en el cuestionario.
b) Estime, con probabilidad 0.95, el intervalo de confianza en el que se encontrará la valoración para esta misma persona.

10. Un test de 5 ítems tiene un coeficiente de fiabilidad de 0.4 y un coeficiente de validez de 0.5.

a) Queremos que su coeficiente de validez alcance el valor de 0.6. ¿Qué longitud debería tener el test?
b) Queremos que su coeficiente de validez alcance el valor de 0.8. ¿Qué longitud debería tener el test?
c) ¿Cuál es máximo valor del coeficiente de validez que se puede alcanzar alargando el test?

11. Un psicólogo dispone de tres pruebas de desorden del pensamiento (T1, T2, y T3), de igual variabilidad, para pronosticar un determinado criterio (esquizofrenia). Los coeficientes de fiabilidad, de validez y número de ítems de cada test son:

	r_{xx}	r_{xy}	ítems
T1	0,3	0,27	20
T2	0,7	0,59	40
T3	0,9	0,60	40

Si los tres tests tuviesen la misma longitud, ¿cuál sería más fiable? ¿cuál, más válido?

12. El coeficiente de determinación de un test es 0.25 y la varianza del criterio es 2.

a) Obtenga el coeficiente de validez y la varianza de los errores de pronóstico.
b) Obtenga, con probabilidad 0.95, la amplitud que tendrá la estimación por intervalo en el criterio para cualquier persona.

13. Asocie cada uno de estos términos a cada una de las frases*: coeficiente de fiabilidad, índice de validez, varianza explicada por un factor, coeficiente de determinación, índice de homogeneidad, saturación.*

a) La correlación de las puntaciones en un ítem con las puntuaciones en el test:
b) La correlación de las puntaciones en un ítem con las puntuaciones en un criterio:
c) La suma de las correlaciones al cuadrado de los ítems con un factor:
d) La correlación de un ítem con un factor:
e) La correlación entre la forma par e impar de un test:
f) La proporción de varianza de un criterio que explica un test:

14. Un test tiene un coeficiente de fiabilidad de 0.6 y de validez de 0.42. Duplicamos el test y conseguimos un coeficiente de validez superior a 0.42. Volvemos a duplicarlo y conseguimos un coeficiente de validez aún mayor.

a) ¿Podrá conseguirse por la vía de sucesivos alargamientos un coeficiente de validez de 0.8? Justifique su respuesta.
b) ¿Podrá conseguirse por la vía de sucesivos alargamientos un coeficiente de fiabilidad de 0.98? Justifique su respuesta.

15. Aplicamos a Juana un test para predecir su rendimiento en un criterio, concluyendo que, con probabilidad 0.99, su puntuación en dicho criterio estará entre 24 y 30. Responda razonadamente a las siguientes cuestiones:

a) ¿Cuál es la estimación puntual que hemos realizado a Juana?
b) ¿Si hubiéramos establecido el intervalo con probabilidad 0.95, su amplitud sería (mayor/menor/igual) a 6?

16. Un Centro de Investigaciones Sociológicas está interesado en evaluar el impacto que los escándalos acaecidos en la vida pública han tenido en la imagen que la sociedad española tiene de la clase política. Para ello, elaboran un cuestionario con seis elementos y la aplican a una muestra de cinco sujetos. Además se les pide a los sujetos que den una valoración personal de la clase política, en una escala de 1 a 20. La tabla recoge las respuestas dadas por los cinco sujetos a las preguntas del cuestionario junto con su valoración de la clase política.

SUJETOS	ITEM						valoración de la clase política
	1	2	3	4	5	6	
1	4	2	3	3	5	4	15
2	4	2	2	5	2	4	10
3	4	3	2	2	4	6	18
4	1	2	1	4	3	4	9
5	2	1	2	1	2	2	13

a) Obtenga e interprete el coeficiente alfa del cuestionario formado por los 6 ítems.
b) ¿Cuál será la validez de este cuestionario si tomamos las valoraciones realizadas por los sujetos como un criterio adecuado? Interprételo.

c) ¿Cuántos elementos tendríamos que añadirle al cuestionario para que su nuevo coeficiente de validez alcanzase un valor de 0,60?

17. Del estudio psicométrico de un test de 4 items, hemos obtenido:

Estadísticos de fiabilidad

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
.433	.414	4

Estadísticos de los elementos

	Media	Desviación típica	N
IT1	2.64	1.295	156
IT2	2.28	1.242	156
IT3	3.45	1.225	156
IT4	3.29	1.158	156

Matriz de correlaciones inter-elementos

	IT1	IT2	IT3	IT4
IT1	1.000	.347	.517	-.133
IT2	.347	1.000	.393	-.060
IT3	.517	.393	1.000	-.165
IT4	-.133	-.060	-.165	1.000

Se ha calculado la matriz de covarianzas y se utiliza en el análisis.

Estadísticos total-elemento

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Correlación múltiple al cuadrado	Alfa de Cronbach si se eleimina el elemento
IT1	9.01	4.942	.410	.294	.169
IT2	9.38	5.295	.374	.183	.220
IT3	8.21	5.106	.428	.328	.160
IT4	8.37	8.672	-.152	.030	.684

Estadísticos de la escala

Media	Varianza	Desviación típica	N de elementos
11.65	8.976	2.996	4

Varianza total explicada

	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción		
Factor	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	1.892	47.302	47.302	1.345	33.631	33.631
2	.963	24.064	71.366			
3	.668	16.707	88.072			
4	.477	11.928	100.000			

Método de extracción: Máxima verosimilitud.

Matriz factorial[a]

	Factor 1
IT1	.671
IT2	.508
IT3	.774
IT4	-.196

Método de extracción: Máxima verosimilitud.

a. 1 factores extraídos. Requeridas 4 iteraciones.

Prueba de la bondad de ajuste

Chi-cuadrado	gl	Sig.
.415	2	.813

RMSEA	Intervalo de confianza 90%
.000	.000-.098

RESPONDA **RAZONADAMENTE** LAS SIGUIENTES PREGUNTAS

a) ¿El test es de rendimiento óptimo?
b) Supongamos que queremos que el test definitivo tenga 3 ítems. ¿Cuál eliminaría si queremos que el test tenga la máxima variabilidad? ¿Cuál sería la variabilidad y consistencia interna del test de 3 ítems resultante?
c) Supongamos que queremos que el test definitivo tenga máxima consistencia y solo dos ítems. ¿Qué dos ítems eliminaría? Calcule e interprete el coeficiente de fiabilidad del test de dos ítems resultante, sabiendo las correlaciones entre ellos.
d) Diga cuanto vale la correlación del ítem 2 con el factor 1.
e) ¿Qué porcentaje de varianza total explica el factor?
f) Según los resultados del análisis factorial, ¿diría que el investigador puede mantener la unidimensionalidad del test?

18. En una muestra de universitarios, que habían superado la selectividad, el coeficiente de validez de un test de conocimientos fue 0.54. ¿Cuál hubiese sido si se hubiese calculado con los datos de todos los aspirantes y no sólo de los que han superado el examen? a) mayor que 0,54; b) menor que 0,54; c) igual (0,54). Razone su respuesta.

19. Si el coeficiente de validez del test es 0.7, la correlación entre las puntuaciones verdaderas entre dicho test y el criterio será: a) 0,7; b) mayor que 0,7; c) menor que 0,7. Razone su respuesta.

20. ¿Son corrrectos los siguientes enunciados sobre la validez de un test?

a) La validez de constructo de un test asegura su validez referida a un criterio. V ( ) F ( ).
b) Para estudiar la validez de contenido de un test no es necesario aplicarlo a una muestra.
V ( ) F ( ).

c) Si el coeficiente de validez de un test vale 0,7 esto significa que el 70% de la variabilidad de las puntuaciones en el criterio se puede pronosticar a partir del test.

V ( ) F ( ).

21. Hemos analizado mediante SPSS un test de 9 Ítems y 5 opciones de respuesta cuyos ítems son los siguientes:

Ítem 1. Me gusta leer libros.
Ítem 2. Entiendo bien las explicaciones de los profesores.
Ítem 3. Me gustan los documentales de "la 2"
Ítem 4. Me gusta ver los telediarios y saber lo que ocurre en el mundo.
Ítem 5. Soy creativo a la hora de inventar juegos.
Ítem 6. Se me dan bien las matemáticas.
Ítem 7. Me gustan las novedades.
Ítem 8. Me atrae la idea de viajar y conocer otras culturas.
Ítem 9. Entiendo todo rápidamente.

Estadísticos total-elemento

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Correlación múltiple al cuadrado	Alfa de Cronbach si se eleimina el elemento
ITEM01	28.32	21.711	.281	.139	.645
ITEM02	27.62	22.300	.482	.477	.597
ITEM03	28.12	22.761	.240	.097	.652
ITEM04	27.65	22.537	.385	.331	.615
ITEM05	28.56	21.576	.408	.208	.607
ITEM06	27.94	22.629	.267	.306	.644
ITEM07	26.80	24.387	.336	.268	.630
ITEM08	26.81	25.381	.165	.232	.657
ITEM09	27.81	21.871	.536	.528	.586

Estadísticos de fiabilidad

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
.654	.674	9

Estadísticos de resumen de los elementos

		Media	Mínimo	Máximo	Rango	Máximo/ mínimo	Varianza	N de elementos
Medias de los elementos	Parte 1	3.149	2.646	3.586	.939	1.355	.172	5[a]
	Parte 2	3.864	3.263	4.404	1.141	1.350	.385	4[b]
	Ambas partes	3.467	2.646	4.404	1.758	1.664	.372	9
Varianzas de los elementos	Parte 1	1.504	.939	2.128	1.189	2.266	.229	5[a]
	Parte 2	1.016	.631	1.706	1.075	2.703	.227	4[b]
	Ambas partes	1.287	.631	2.128	1.497	3.372	.266	9
Correlaciones inter-elementos	Parte 1	.193	.087	.394	.308	4.542	.008	5[a]
	Parte 2	.154	-.143	.467	.609	-3.273	.052	4[b]
	Ambas partes	.187	-.143	.633	.776	-4.442	.027	9

Se ha calculado la matriz de covarianzas y se utiliza en el análisis.

a. Los elementos son: ITEM01, ITEM02, ITEM03, ITEM04, ITEM05.

b. Los elementos son: ITEM06, ITEM07, ITEM08, ITEM09.

Estadísticos de fiabilidad

Alfa de Cronbach	Parte 1	Valor	.531
		N de elementos	5[a]
	Parte 2	Valor	.403
		N de elementos	4[b]
	N total de elementos		9
Correlación entre formas			.502
Coeficiente de Spearman-Brown	Longitud igual		.669
	Longitud desigual		.671
Dos mitades de Guttman			.634

a. Los elementos son: ITEM01, ITEM02, ITEM03, ITEM04, ITEM05.

b. Los elementos son: ITEM06, ITEM07, ITEM08, ITEM09.

Varianza total explicada

Factor	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2.641	29.349	29.349	2.108	23.425	23.425	1.915	21.276	21.276
2	1.755	19.498	48.847	1.205	13.388	36.812	1.398	15.536	36.812
3	1.059	11.772	60.619						
4	.840	9.339	69.958						
5	.738	8.203	78.160						
6	.605	6.724	84.884						
7	.569	6.320	91.204						
8	.484	5.379	96.583						
9	.307	3.417	100.000						

Método de extracción: Máxima verosimilitud.

Matriz factorial[a]

	Factor	
	1	2
ITEM01	.255	.290
ITEM02	.744	-.217
ITEM03	.222	.054
ITEM04	.379	.684
ITEM05	.469	.058
ITEM06	.512	-.356
ITEM07	.396	.348
ITEM08	.080	.575
ITEM09	.807	-.146

Método de extracción: Máxima verosimilitud.

a. 2 factores extraídos. Requeridas 5 iteraciones.

Matriz de factores rotados[a]

	Factor	
	1	2
ITEM01	.128	.364
ITEM02	.771	.079
ITEM03	.185	.133
ITEM04	.094	.776
ITEM05	.413	.230
ITEM06	.608	-.138
ITEM07	.237	.471
ITEM08	-.142	.563
ITEM09	.803	.168

Método de extracción: Máxima verosimilitud.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 3 iteraciones.

Prueba de Bondad de ajuste modelo de dos factores:

Prueba de la bondad de ajuste

Chi-cuadrado	gl	Sig.
14.327	19	.764

RMSEA	Intervalo de confianza 90%
.000	.000-.068

Prueba de Bondad de ajuste modelo de un factor:

Prueba de la bondad de ajuste

Chi-cuadrado	gl	Sig.
64.403	27	.000

RMSEA	Intervalo de confianza 90%
.124	.087-.161

Responda a las siguientes preguntas **RAZONADAMENTE:**

a) Diga si el test es de rendimiento típico o de rendimiento óptimo.
b) Diga cuál es la varianza explicada por el segundo factor.
c) ¿Puede decirse que el test es unidimensional?
d) Interprete el significado de los factores.
e) La primera mitad del test está formada por los ítems_____ , y su consistencia interna es_____.
f) Asumiendo que ambas mitades son paralelas obtenga e interprete el coeficiente de fiabilidad de cualquiera de ellas
g) Si tuviera que eliminar un ítem diga qué ítem eliminaría y por qué _____ .
h) Atendiendo a la columna "Alpha if item deleted" diga cuales son los dos ítems que más correlacionan con el total del test.

22. A un grupo normativo de 100 sujetos se le ha aplicado un test (X) formado por 4 ítems y se le ha medido en un criterio (Y), obteniéndose la siguiente matriz de correlaciones. Se indica también la varianza de cada variable.

	Item 1	Item 2	Item 3	Item 4	X	Y
Item 1	1,00					
Item 2	0,70	1,00				
Item 3	0,00	0,25	1,00			
Item 4	-0,44	-0,31	-,31	1,00		
X	0,71	0,87	0,5	-0,15	1,00	
Y	0,75	0,68	-0,08	0,14	0,76	1,00
Varianza	0,30	0,27	0,27	0,17	1,06	25,9

a) Diga cuáles son: 1) el ítem que más contribuye a la consistencia interna del test, 2) el ítem que más contribuye a la validez del test. Razone sus respuestas y, en caso de necesidad, realice los cálculos oportunos.
b) Obtenga e interprete un indicador de la consistencia interna del test.
c) Obtenga la amplitud que tendría el intervalo para estimar la puntuación en el criterio de cualquier persona que hiciera el test, si dicho intervalo lo establecemos con un nivel de significación de 0.05.
d) Sabiendo que el coeficiente de fiabilidad del test de 4 ítems es 0.80, obtenga e interprete el coeficiente de validez que tendría el test si le añadimos 12 ítems paralelos a los que ya tiene.

23. Un test de responsabilidad consta de 25 ítems. Su coeficiente de fiabilidad test-retest fue 0.82, su media 30 y su varianza empírica 16. La correlación entre el test y un criterio externo fue 0.40, siendo la varianza del criterio igual a 20 y su media 50.

a) Una persona obtiene en el test de responsabilidad una puntuación que se encuentra dos desviaciones típicas por debajo de la media. Obtenga, con probabilidad 0.95, entre qué valores estimamos que se encontrará su puntuación directa en el criterio.
b) Obtenga e interprete la proporción de varianza del criterio que podemos pronosticar con el test si lo alargamos 3 veces, es decir, si le añadimos dos formas paralelas.

24. A continuación aparecen distintas partes de una salida de SPSS para el análisis de la fiabilidad de 6 de los 36 ítems del test de matrices progresivas de Raven en una muestra de 1800 sujetos. En concreto, se analizaron los ítems que ocupaban las posiciones 10ª, 15ª, 20ª, 25ª, 30ª y 35ª del test.

Matriz factorial[a]

	Factor
	1
raven10	.484
raven15	.245
raven20	.260
raven25	.358
raven30	.308
raven35	.245

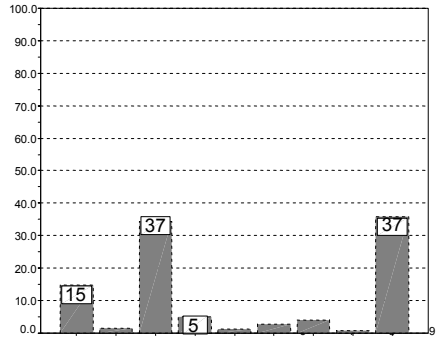
Método de extracción: Máxima verosimilitud.
a. 1 factores extraídos. Requeridas 3 iteraciones.

Prueba de la bondad de ajuste

Chi-cuadrado	gl	Sig.
15.869	9	.070

RMSEA	Intervalo de confianza 90%
.021	.000-.037

ANÁLISIS DE LAS OPCIONES (ÍTEM 35)



Estadísticos de fiabilidad

Alfa de Cronbach	Parte 1	Valor	.258
		N de elementos	3[a]
	Parte 2	Valor	.255
		N de elementos	3[b]
	N total de elementos		6
Correlación entre formas			.229
Coeficiente de Spearman-Brown	Longitud igual		.373
	Longitud desigual		.373
Dos mitades de Guttman			.370

a. Los elementos son: raven10, raven15, raven20.

b. Los elementos son: raven25, raven30, raven35.

Estadísticos de los elementos

	Media	Desviación típica	N
raven10	.82	.381	1800
raven15	.78	.417	1800
raven20	.70	.456	1800
raven25	.61	.488	1800
raven30	.58	.494	1800
raven35	.37	.484	1800

Estadísticos de resumen de los elementos

		Media	Mínimo	Máximo	Rango	Máximo/ mínimo	Varianza	N de elementos
Medias de los elementos	Parte 1	.769	.704	.824	.120	1.170	.004	3[a]
	Parte 2	.519	.373	.608	.235	1.629	.016	3[b]
	Ambas partes	.644	.373	.824	.451	2.208	.027	6
Varianzas de los elementos	Parte 1	.176	.145	.208	.064	1.439	.001	3[a]
	Parte 2	.239	.234	.244	.010	1.043	.000	3[b]
	Ambas partes	.207	.145	.244	.099	1.687	.002	6
Correlaciones inter-elementos	Parte 1	.108	.056	.156	.100	2.777	.002	3[a]
	Parte 2	.102	.080	.115	.035	1.429	.000	3[b]
	Ambas partes	.098	.036	.174	.138	4.830	.002	6

Se ha calculado la matriz de covarianzas y se utiliza en el análisis.

a. Los elementos son: raven10, raven15, raven20.

b. Los elementos son: raven25, raven30, raven35.

Estadísticos total-elemento

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-tot al corregida	Correlación múltiple al cuadrado	Alfa de Cronbach si se eleimina el elemento
raven10	3.04	1.449	.269	.076	.302
raven15	3.09	1.529	.133	.029	.376
raven20	3.16	1.456	.159	.030	.361
raven25	3.26	1.365	.207	.052	.330
raven30	3.29	1.369	.196	.040	.338
raven35	3.49	1.434	.148	.028	.370

Estadísticos de la escala

	Media	Varianza	Desviación típica	N de elementos
Parte 1	2.31	.636	.797	3[a]
Parte 2	1.56	.864	.929	3[b]
Ambas partes	3.86	1.840	1.356	6

a. Los elementos son: raven10, raven15, raven20.

b. Los elementos son: raven25, raven30, raven35.

En la figura se representan los porcentajes de elección de las opciones de respuesta para el ítem 35, cuya opción correcta es la 3 (recuerde que el test se ha aplicado a 1800 sujetos y que cada ítem tiene 8 opciones de respuesta).

RESPONDA **RAZONADAMENTE** A LAS SIGUIENTES PREGUNTAS:

a) Atendiendo a los datos del análisis de fiabilidad y a la figura ¿Cuál es el índice de dificultad del ítem 35 considerando las omisiones como datos perdidos? ¿Qué número de personas omiten este ítem?
b) Observando la frecuencia de elección de las opciones del ítem 35 ¿se incumple alguno de los supuestos de la fórmula de corrección del azar? ¿Cuál?
c) Considerando que el patrón de omisiones es parecido para todos los ítems, ¿cree usted que en general habrá muchas diferencias entre las puntuaciones directas y las correspondientes corregidas?
d) Obtenga e interprete el coeficiente de fiabilidad del test completo de 36 ítems.
e) De las 2 partes del test que se consideran en el análisis, diga cuáles son los ítems que forman la mitad más consistente.
f) Algunos autores obtienen datos que indican que sobre el rendimiento de los 36 ítems del Raven subyacen dos dimensiones cognitivas. ¿Los datos que se muestran para los 6 ítems analizados van en esta línea?
g) El ítem que más contribuye al primer factor es el…….. ya que su correlación con dicho factor es …….
h) ¿Cuál es el porcentaje de varianza total explicado por el primer factor antes de rotar?
i) ¿Crees que las dos mitades que hemos formado son formas paralelas?

25. A continuación aparecen distintas partes de una salida de SPSS correspondientes a 8 ítems, en una muestra de 102 sujetos.

Estadísticos de fiabilidad

Alfa de Cronbach	Alfa de Cronbach basada en los elementos tipificados	N de elementos
.592	.598	8

Estadísticos de los elementos

	Media	Desviación típica	N
item1	1.66	1.278	102
item2	1.98	1.134	102
item3	1.54	1.224	102
item4	1.68	1.055	102
item5	1.72	1.093	102
item6	2.60	1.065	102
item7	2.22	1.059	102
item8	2.65	1.157	102

Estadísticos total-elemento

	Media de la escala si se elimina el elemento	Varianza de la escala si se elimina el elemento	Correlación elemento-total corregida	Correlación múltiple al cuadrado	Alfa de Cronbach si se eleimina el elemento
item1	14.37	17.444	.219	.368	.586
item2	14.05	16.918	.344	.167	.544
item3	14.49	17.401	.246	.188	.575
item4	14.35	18.627	.184	.095	.590
item5	14.31	16.811	.380	.303	.533
item6	13.43	17.456	.318	.132	.553
item7	13.81	16.470	.445	.338	.515
item8	13.38	17.684	.246	.336	.574

Estadísticos de fiabilidad

Alfa de Cronbach	Parte 1	Valor	.216
		N de elementos	4[a]
	Parte 2	Valor	.448
		N de elementos	4[b]
	N total de elementos		8
Correlación entre formas			.550
Coeficiente de Spearman-Brown	Longitud igual		.710
	Longitud desigual		.710
Dos mitades de Guttman			.710

a. Los elementos son: item1, item2, item3, item4.

b. Los elementos son: item5, item6, item7, item8.

Varianza total explicada

	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
Factor	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	2.199	27.490	27.490	1.373	17.163	17.163	1.534	19.178	19.178
2	1.670	20.873	48.363	1.542	19.275	36.438	1.381	17.260	36.438
3	1.000	12.500	60.862						
4	.867	10.834	71.696						
5	.784	9.805	81.501						
6	.586	7.322	88.823						
7	.498	6.220	95.043						
8	.397	4.957	100.000						

Método de extracción: Máxima verosimilitud.

Matriz factorial[a]

	Factor	
	1	2
item1	.973	-.004
item2	.194	.385
item3	-.052	.447
item4	-.003	.328
item5	-.048	.658
item6	.188	.306
item7	-.029	.745
item8	.589	.066

Método de extracción: Máxima verosimilitud.

a. 2 factores extraídos. Requeridas 36 iteraciones.

Matriz de factores rotados[a]

	Factor	
	1	2
item1	-.045	.972
item2	.376	.210
item3	.449	-.034
item4	.327	.010
item5	.660	-.021
item6	.298	.201
item7	.746	.003
item8	.041	.591

Método de extracción: Máxima verosimilitud.
Método de rotación: Normalización Varimax con Kaiser.

a. La rotación ha convergido en 3 iteraciones.

Prueba de Bondad de ajuste modelo de dos factores:

Prueba de la bondad de ajuste

Chi-cuadrado	gl	Sig.
7.648	13	.866

RMSEA	Intervalo de confianza 90%
.000	.000-.057

Prueba de Bondad de ajuste modelo de un factor:

Prueba de la bondad de ajuste

Chi-cuadrado	gl	Sig.
55.004	20	.000

RMSEA	Intervalo de confianza 90%
.136	.096-.178

Responda **razonadamente** a las siguientes preguntas:a) Asumiendo que las dos mitades son formas paralelas, obtenga e **interprete** el coeficiente de fiabilidad del

a.1) test completo de 8 ítems

a.2) subtest formado por los ítems 5, 6,7 y 8

b) Para maximizar la varianza del test habría que eliminar el ítem _____, ya que …

c) El índice de homogeneidad corregido del ítem 2 es _____. Con un nivel de confianza del 95%, ¿considera que la relación entre ese ítem y la puntuación en el resto del test es significativamente distinta de cero?

d) Para maximizar la consistencia interna del test habría que eliminar el ítem _____, ya que …

e) Atendiendo a toda la información de las tablas, ¿convendría eliminaría algún ítem? En caso de respuesta afirmativa, diga qué ítem o ítems convendría eliminar.

f) Obtenga e **interprete** la comunalidad del ítem 1.

g) Deberían extraerse ____ factores, ya que …

h) Para interpretar el significado del factor 2 utilizaría los ítems _____, ya que …

i) La proporción de varianza total explicada por el factor I no rotado es _____, y por el factor II rotado es _____.

SOLUCIONES

1. a) Obtener su coeficiente de validez.
b) Aportar información sobre su validez de contenido.
c) Aportar datos sobre su validez factorial. Cabe pensar que si el test es válido, sature en el mismo factor que los otros tests que miden constructos relacionados.

2. El coeficiente de validez viene determinado por varios factores:
- La fiabilidad del test y del criterio.
- La longitud de ambos.
- La variabilidad del grupo normativo en ambos.
- La auténtica relación entre el test y el criterio.

3. No. La fiabilidad del test podemos incrementarla hasta que llegue a su tope de 1. En este caso, la validez máxima que puede alcanzar este test completamente fiable sería la raíz cuadrada del coeficiente de fiabilidad del criterio, que sería igual a 0.77.

4. No estaría totalmente justificada la desestimación del cuestionario, dado que, por tener tan pocos ítems, obtendríamos un incremento apreciable en sus cualidades psicométricas al incrementar su longitud.

5. a) $r_{xy} = 0.98$
b) $S_{y\text{-}y'} = 0.45$

6. a) Sí. Por ejemplo, cuando el criterio no tenga que ver con lo que el test mida.
b) No, ya que $r_{xy} \leq \sqrt{0.25} = 0.5$.

7. a) Parece que sí, dado que los dos ítems de aptitud verbal obtienen saturaciones altas en el Factor II, mientras que los dos ítems de aptitud numérica obtienen saturaciones altas en el Factor I.
b) El porcentaje de varianza explicado por el factor I será (1.77)(100)/4 = 44.

8. El Factor I podría definirse como un "producto bueno".El Factor II como un "producto barato" y el Factor III como un "producto bonito".

9. a) Y' = 9.475
b) $L_i = 3.97$ $L_s = 14.98$

10. a) n = 2.04 ≅ 2, debería estar formado por 2 formas paralelas del test inicial; es decir, por 10 ítems.
b) n = -64, no se puede alcanzar la validez de 0.8 mediante el incremento de la longitud del test.
c) El máximo valor del coeficiente de validez obtenible por alargamiento del test es $R_{xy} \leq r_{xy} / \sqrt{r_{xx}} = 0.5 / \sqrt{0.4} = 0.79$.

11. Si los tres tests tuviesen la misma longitud el más fiable sería el tercero, ya que si alargásemos el Test 1 hasta que tuviese 40 ítems su coeficiente de fiabilidad valdría 0,46. El más válido sería también el Test 3, ya que al alargar el Test 1 su coeficiente de validez toma el valor de 0,33.

12. $r_{x\,y}^2 = 0.25$; $S_y^2 = 2$
Coeficiente de validez: 0.5
Varianza de los errores de pronóstico: 1.5
Amplitud del intervalo: 4.80

13. a) Indice de homogeneidad
b) Índice de validez
c) varianza explicada por un factor
d) Saturación
e) Coeficiente de fiabilidad
f) Coeficiente de determinación

14. a) 0.8 es mayor que el tope máximo alcanzable (0.54 = 0.42/√0.6) alargando el test, luego NO se puede alcanzar el valor 0.8.
b) Cualquier valor menor que 1 se puede alcanzar alargando el test. Luego, SI.

15. a) Y' = (30 + 24)/ 2 = 27.
b) Menor. Con probabilidad 0.99, A= (2)2.57S. Con probabilidad 0.95, A = (2)1.96S.

16. a) Las varianzas de los ítems son: 1.6, 0.4, 0.4, 2, 1.36 y 1.6
La varianza del test es 17.76
El coeficiente alfa es (6/5)(1 – (7.36/17.67)) = 0.70. Alta consistencia, pues el test es corto.
b) El coeficiente de validez es 0.43. El 18% de la varianza del criterio depende del test.
c) La correlación par e impar es 0.33. El coeficiente de fiabilidad del test (dos mitades) es 0.50. El número de formas paralelas necesarias para alcanzar la validez 0.6 es 36.73. El test deberá tener 36.73x6= 220.38 ítems, por lo que deberemos añadir 220.38 – 6 = 214.38 ó 215 ítems.

17. a) No, pues las medias son mayores que 1.
b) Eliminaríamos el ítem 4. La varianza del test resultante sería 8.672. Su coeficiente alfa sería 0.684.
c) Hay que quitar los ítems 2 y 4. El test formado por los ítems 1 y 3 tendría un coeficiente de fiabilidad de 0.682 (dos mitades), pues la correlación entre la mitad par e impar es 0.517 y aplicando Spearman-Brown, el coeficiente de fiabilidad es 0.682.
d) 0.508
e) 33.63%
f) Claramente no. Aunque el RMSEA indica un buen ajuste del modelo de un factor, el peso del ítem 4 es *negativo*.

18. Sería mayor que 0.54, pues la muestra de todos los aspirantes (los que han aprobado la selectividad y los que no) tiene una mayor variabilidad y por lo tanto cabe esperar un mayor coeficiente de validez.

19. La correlación entre las puntaciones verdaderas del test y criterio es mayor o igual que el coeficiente de validez (véase apartado 4.3).

20.
a) F
b) V
c) F

21.
a) Típico (las medias de los ítems son mayores de 1).
b) 1.398 (el % de varianza explicada sería 15.536).
c) No. Hemos retenido dos factores. El modelo de un factor no ajusta bien a los datos. El estadístico de contraste muestra que con un nivel de confianza del 95%, podemos decir que algún residual es distinto de cero. Además el RMSEA es mayor que 0.08. Por el contrario, para el modelo de dos factores, los indicadores de ajuste muestran valores aceptables (el RMSEA nos indica que el modelo muestra buen ajuste a los datos ya que su valor es menor que 0.05).
d) Para el factor 1, se utilizarían los ítems 2, 5, 6 y 9. Podríamos ponerle la etiqueta de "*Percepción de la propia capacidad*". Para el factor 2, se utilizarían los ítems 1, 4, 7 y 8. La etiqueta, atendiendo al contenido común de esos ítems, podría ser "*Curiosidad intelectual*".
e) Primera mitad: ítems 1 a 5. Su alfa es 0.531.
f) El coeficiente de fiabilidad del test de 10 ítems es 0.671 = 2r/(1+r), siendo *r* la correlación entre las dos partes y el coeficiente de fiabilidad de cada una. Despejando, r = 0.505.
g) El ítem 8 que tiene el menor índice de homogeneidad corregida (HC = 0.165). Además, al eliminarlo aumenta alfa desde 0.654 a 0.657
h) Los dos que tengan menores valores en esa columna: ítems 2 y 9.

22. El ítem 2, pues tiene la mayor correlación con el test (0.87). El ítem 4, pues tiene la mayor diferencia V-H (0.29).
b) El coeficiente alfa vale 0.06. El test no tiene consistencia.
c) 12.97
d) El coeficiente de validez del test alargado cuatro veces es 0.82. Este coeficiente de validez es muy alto. El test predice muy bien el criterio. El 67.24% de la varianza del criterio puede explicarse por las puntuaciones en el test

23. a) (38.27, 54.43).
b) El coeficiente de validez del test alargado es 0.43. La proporción pedida es 0.18. Es un coeficiente de validez medio.

24.
a) 0.59, pues 0.37/(1-0.37) = 0.587. Lo omiten 666 = (0.37)1800.
b) Las alternativas no son igualmente elegidas.
c) Las diferencias entre puntuaciones y puntuaciones corregidas serán pequeñas, pues el número de errores en los ítems es bajo y el número de opciones en cada ítem es alto.
d) El coeficiente de fiabilidad por el método de las dos mitades es 0.373. Alargando el test 6 veces, resulta un test con coeficiente de fiabilidad dos mitades de 0.781. El 78% de la varianza observada se debe a la varianza de los niveles de rasgo. Es una fiabilidad aceptable para un test de esa longitud.

e) La mitad más consistente es la parte 1 (ítems 10, 15 y 20). Su alfa es 0.258.
f) No, pues el ajuste del modelo de un factor es bueno según ambos indicadores de ajuste. Podemos mantener que el modelo se ajusta a los datos con un nivel de confianza del 95% y además el RMSEA indica un buen ajuste (RMSEA < 0.05).
g) Ítem 10, pues la correlación es 0.484.
h) 10.756 (suma de saturaciones al cuadrado dividido por 6 y multiplicado por 100).
i) Claramente no, las dos mitades difieren en media puesto que los ítems están ordenados por dificultad.

25.
a1) El coeficiente de fiabilidad del test es 0.71. El 71% de la varianza de las puntuaciones observadas corresponde a variabilidad de las puntuaciones verdaderas (y el 29% al error de medida).
a2) 0.550, pues es la correlación entre las dos partes.
b) Para maximizar la varianza del test habría que eliminar el ítem 4, ya que al eliminarlo la varianza del test de 7 ítems alcanza el valor más alto (18.627).
c) El índice de homogeneidad corregido del ítem 2 es 0.344. Si sería significativo, pues $.344\sqrt{102} = 3.474 > 1.96$.
d) Para maximizar la consistencia interna del test habría que eliminar el ítem 4, ya que, al eliminarlo, se conseguiría que el test de 7 ítems tenga el mayor alfa (0.590).
e) El ítem 4, por que prácticamente no cambia el coeficiente alfa del test al eliminar ese ítem. Además, la puntuación en ese ítem no correlacióna significativamente con la puntuación en el resto del test ($.184\sqrt{102} = 1.858 < 1.96$)
f) La comunalidad del ítem 1 es $(-0.045)^2+(0.972)^2 = 0.947$. El 95% de la varianza del ítem 1 está explicado por los dos factores.
g) Deberían extraerse 2 factores, ya que los indicadores de ajuste muestran que el modelo de un factor no se ajusta a los datos, mientras que el modelo de dos factores sí.
h) Para interpretar el significado del factor 2 utilizaría los ítems 1 y 8, ya que son los que tienen saturaciones más altas en ese factor, en la matriz rotada.
i) La proporción de varianza total explicada por el factor I no rotado es 1.534/8 = 0.19, y por el factor II rotado es 1.381/8 = 0.17.

TEMA V: BAREMACIÓN DE UN TEST

1.- INTRODUCCIÓN

La puntuación directa de una persona en un test no es directamente interpretable si no la referimos a los contenidos incluidos en el test o al rendimiento de las restantes personas que comparten el grupo normativo. Nosotros centramos en este segundo sentido el tema de la interpretación de una puntuación directa en un cuestionario, para lo cual es necesario tratar el tema de la obtención de baremos para comparar esta puntuación con las que obtienen las personas que han formado el grupo normativo. De una u otra forma, los baremos consisten en asignar a cada posible puntuación directa un valor numérico (en una determinada escala) que informa sobre la posición que ocupa la puntuación directa (y por tanto la persona que la obtiene) en relación con los que obtienen las personas que integran el grupo normativo donde se bareman las pruebas.

Entre las múltiples formas de baremar un test, destacamos las siguientes:

- Baremos cronológicos: Edad Mental y Cociente Intelectual.

- Centiles o percentiles.

- Puntuaciones típicas: estándares, normalizadas, escalas T y D, estaninos o eneatipos.

Lo más usual en las pruebas comercializadas es encontrarse baremos realizados en escala de centiles ó estaninos.

2.- BAREMOS CRONOLÓGICOS

Para rasgos psicológicos que evolucionan con la edad (sobre todo de tipo intelectual) tiene sentido comparar la puntuación de un sujeto con las que obtienen los de su misma edad y los de edades diferentes. Esto se puede realizar mediante dos tipos diferentes de baremos: las Edades Mentales (EM) y los Cocientes Intelectuales (CI).

Supongamos que aplicamos un test de Inteligencia de dificultad progresiva a diferentes grupos de edad (niños entre 5 y 14 años), y que obtenemos las puntuaciones medias de cada grupo de edad en la prueba, siendo las que siguen:

Edad:	5	6	7	8	9	10	11	12	13	14
Media:	6	8	9	11	14	15	18	22	24	27

Hemos realizado una correspondencia entre las edades y puntuaciones medias que nos va permitir obtener la EM de cualquier niño al que apliquemos el test. Por ejemplo, si un niño obtiene el test una puntuación directa de 14 puntos, le asignamos una EM de 9 años, independientemente de su edad cronológica real, ya que esa puntuación es la media que obtienen los niños de 9 años.

El Cociente Intelectual (CI) se denomina así (y no coeficiente, como es usual escuchar en determinados ámbitos) porque es el resultado de dividir la edad mental (EM) entre la edad cronológica (EC) del sujeto; para evitar decimales el resultado se multiplica por 100, de tal manera que se puede obtener a partir de la fórmula:

$$CU = \frac{EM}{EC}100$$

Por ejemplo, en el ejemplo anterior, si un niño de 10 años obtiene una puntuación directa de 18 puntos, diremos que su EM es de 11 años, y que su CI es:

$$CU = \frac{11}{10}100 = 110$$

Podemos observar que si la EM de un sujeto coincide exactamente con su EC, el CI es igual a 100, e indicará que este sujeto obtiene exactamente la puntuación media de su grupo de edad. Si el CI supera el valor de 100 significará que el sujeto tiene una Inteligencia superior al promedio de su edad, mientras que si su CI es inferior a 100, significa que el sujeto tiene una inteligencia inferior a la media de su grupo de edad. Usualmente, Cocientes Intelectuales inferiores a 70 indican problemas importantes (deficiencias) de tipo cognitivo, mientras que Cocientes Intelectuales superiores a 140 indican excepcionalidad intelectual.

3.- CENTILES O PERCENTILES

Los centiles, como recordaremos, representan medidas de posición en un distribución de frecuencias. Los baremos centiles consisten en asignar a cada posible puntuación directa un valor (en una escala de 1 a 100) que se denomina centil (o percentil) y que indican el porcentaje de sujetos del grupo normativo que obtienen puntuaciones iguales o inferiores a las correspondientes directas. Así, si un sujeto obtiene en un cuestionario de autoritarismo una puntuación de 20 puntos, poco sabemos sobre su nivel de autoritarismo, pero si sabemos que a esa puntuación le corresponde el centil 95 , ya conocemos que este sujeto supera en ese rasgo al 95% de los sujetos utilizados para baremar el test; si el grupo normativo fuese una muestra

representativa de la población general, podríamos inferir que esta persona supera en autoritarismo al 95% de las personas, y que sólo un 5% de personas son más autoritarias que él.

El modo de cálculo del centil asociado a una puntuación se resume en los siguientes pasos:

1) Disponer en una columna, ordenadas de mayor a menor o de menor a mayor, las posibles puntuaciones directas (X_i) que se puedan obtener en el test.

2) Asignar a cada puntuación su frecuencia (f_i), es decir, el nº de sujetos del grupo normativo que la han obtenido.

3) Disponer una tercera columna de frecuencias acumuladas (F_i).

4) Para cada valor de F_i, obtener el valor C_i= (100) F_i/N, siendo Ci el centil asignado a la puntuación directa X_i, F_i la frecuencia acumulada correspondiente a X_i y N el número total de sujetos que forman el grupo normativo.

Ejemplo:

Supongamos que aplicamos un cuestionario de conocimientos en el manejo de ordenadores a un grupo de 200 universitarios y que las puntuaciones directas obtenidas (X) y los sujetos que obtuvieron cada una de ellas (f) son las siguientes:

X:	28	27	26	25	24	23	22	21	20	19	18
f:	2	4	21	32	45	37	22	18	12	6	1

A partir de estos datos, los centiles correspondientes a cada puntuación directa, se obtienen de la siguiente forma:

X	f	F	Centiles C = (100)F/200
28	2	200	100
27	4	198	99
26	21	194	97
25	32	173	86’5
24	45	141	70’5
23	37	96	48
22	22	59	29’5
21	18	37	18’5
20	12	19	9’5
19	6	7	3’5
18	1	1	0’5

Así, si un sujeto obtiene una puntuación directa de 20 puntos en el cuestionario, diremos que supera en conocimientos informáticos al 9,5% de los sujetos universitarios, mientras que más del 90% de los alumnos universitarios tienen mayor conocimiento en el manejo de ordenadores que la persona evaluada.

4.- PUNTUACIONES TÍPICAS

En Análisis de Datos se vio el significado y proceso de cálculo de las puntuaciones típicas (Z_x) asociadas a unas puntuaciones directas determinadas. En este apartado vamos a encontrar una clara aplicación de estas puntuaciones, y de otras que se derivan de éstas, para baremar un cuestionario; vamos a diferenciar además entre baremos típicos estándares y baremos normalizados.

4.1.- PUNTUACIONES TÍPICAS ESTÁNDARES

Como sabemos, una puntuación típica Z_i se obtiene haciendo:

$$Z_i = \frac{X_i - \overline{X}}{S_X}$$

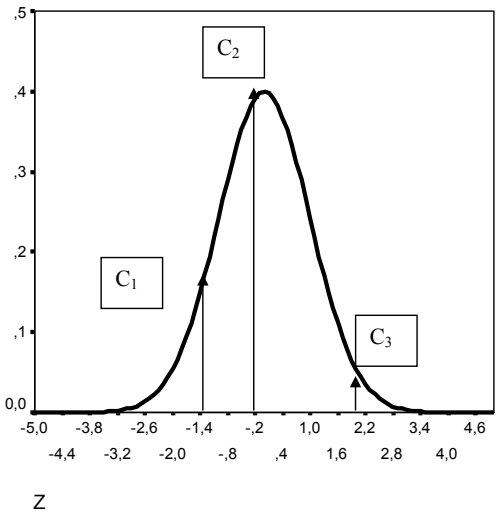
puede ser positiva o negativa, e indica el nº de desviaciones típicas que se aleja de la media una determinada puntuación directa.

Así, conociendo la puntuación típica de un sujeto en un test y la variabilidad del grupo normativo, podemos interpretar el nivel de rasgo del sujeto (atendiendo a la cuantía y signo de su puntuación Z_i) en comparación con los niveles de los restantes sujetos. Por ejemplo, una puntuación típica de -2,33 indica que es un sujeto cuya puntuación se encuentra 2,33 desviaciones típicas por debajo de la media.

4.2.- PUNTUACIONES TÍPICAS NORMALIZADAS

Cuando se puede asumir (o se comprueba) que las puntuaciones de un grupo normativo en un test siguen una distribución normal, un centil concreto C_i dividido entre 100 indica el área de la curva normal que queda por debajo de la puntuación correspondiente.

Por ejemplo, observando la curva normal de la figura, podemos constatar que el C_1 es aproximadamente el centil 10, y deja por debajo un área de 0,10 de la curva normal; el C_2 es el centil 42, y deja por debajo una proporción de 0,42 del área de la curva normal; el C_3 es aproximadamente el centil 95, y deja por debajo de sí un área de 0,95 de la curva normal.



Pues bien, conociendo la proporción que queda por debajo de un punto dado de la distribución, podemos utilizar la tabla de la curva normal para obtener sin cálculos la puntuación típica asociada (Z_n), que se denominará puntuación típica normalizada. Indicará el número de desviaciones típicas que una puntuación se encuentra por encima (si es positiva) o por debajo (si es negativa) de la media en una distribución normal.

Por ejemplo, las puntuaciones típicas normalizadas asociadas a los centiles 1, 26, 57 y 97 son:

Centil	Centil/100	Z_n
1	0,01	-2,33
26	0,26	-0,64
57	0,57	0,18
97	0,97	1,88

Si no se puede asumir racionalmente o no se puede comprobar que las puntuaciones siguen una distribución normal, no se puede hacer uso de las tablas de la curva normal para obtener las Z_n. Sí podrían calcularse las puntuaciones típicas estándares Z_x, ya que no asumimos ningún supuesto sobre la distribución de los datos. Si los datos de una muestra se ajustan a la normal, entonces cada Z_x de una persona es similar a su Z_n.

4.3.- PUNTUACIONES TÍPICAS DERIVADAS

Las puntuaciones típicas (estándares y normalizadas) tienen dos dificultades formales para su interpretación: la posibilidad de asumir valores no enteros y negativos. Con objeto de superar estas pequeñas dificultades, se han propuesto otros baremos, que no son más que una transformación lineal de las puntuaciones típicas, con lo que no se alteran las propiedades de la escala típica. Estas puntuaciones se denominan **escalas típicas derivadas** (si el objeto de la transformación lineal es una puntuación típica estándar) o **escalas típicas derivadas normalizadas** (si suponen la transformación lineal de una puntuación típica normalizada), siendo las principales las denominadas como escala T, escala D y estaninos (o eneatipos):

Escala	Derivada	Derivada y normalizada
Escala T	$T_i = 50+(10)Z_i$	$T_{ni} = 50+(10)Z_{ni}$
Escala D	$D_i = 50+(20)Z_i$	$D_{ni} = 50+(20)Z_{ni}$
Estaninos		$E_{ni} = 5 + 2\ Z_{ni}$

En definitiva, las *puntuaciones T* representan una escala con media 50 y desviación típica 10. Así, una puntuación T = 78 significa que la persona obtiene una puntuación Z_i = 2.8, es decir, 2.8 desviaciones típicas por encima de la media del grupo normativo.

Las *puntuaciones D* suponen una escala con media 50 y desviación típica 20. Por ejemplo, una puntuación D = 35 indica que la persona obtuvo una puntuación Z_i = -.75, o lo que es lo mismo, una puntuación que se encuentra .75 desviaciones típicas por debajo de la media del grupo normativo donde se barema el test.

Los *estaninos* representan otra escala con media 5 y desviación típica 2. Una persona que obtenga el estanino 8 en un test de aptitud espacial indicará que se encuentra 1.5 desviaciones típicas por encima de la media del grupo normativo.

Consideremos un caso de baremación de una misma puntuación en diferentes escalas. Por ejemplo, a un sujeto que obtiene una puntuación directa de 30 puntos en un test de aptitud mecánica con media de 38 puntos y desviación típica 4, le podemos asignar puntuaciones en los siguientes baremos:

Puntuación típica: -2
Escala T: 30
Escala D: 10
Estanino: 1

Todas estas puntuaciones en escalas o baremos diferentes indican lo mismo: que es un sujeto que se encuentra dos desviaciones típicas por debajo de la media de grupo normativo en aptitud mecánica.

La interpretación de cada una de las escalas típicas derivadas normalizadas sigue la misma lógica que su correspondiente escala típica derivada sin normalizar, haciendo siempre la salvedad de que la interpretación hay que referirla a una distribución normal.

EJERCICIOS

1. ¿Cuál es el objetivo de la construcción de baremos?

2. Un grupo de 200 personas obtuvo en un test de inteligencia una media de 14.78 puntos y una desviación típica de 3.34. La siguiente tabla recoge la distribución de frecuencias de las puntuaciones obtenidas por los sujetos en el test:

X	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
f	2	3	4	11	17	15	23	20	19	23	22	15	8	8	4	5	1

a) Calcule los centiles correspondientes a cada una de las puntuaciones directas.
b) Calcule las puntuaciones típicas, puntuaciones T y D que corresponden a sujetos con puntuaciones directas de 10 y 21 puntos.
c) Suponiendo que la distribución se adapta a la distribución de la curva normal, que puntuaciones típicas normalizadas y en las escalas derivadas (T, D y E) corresponderían a esos mismos sujetos.

3. La media de un test es 45 y la desviación típica 10. Sabemos que a la puntuación directa de 40 le corresponde el centil 21, y que en las tablas de la curva normal la puntuación típica -0,8 deja por debajo de sí la probabilidad de 0,21. Calcule el valor asociado a la puntuación directa de 40 en las siguientes escalas:

a) Centil.
b) Típica normalizada (Z_n).
c) Escala D no normalizada.
d) Estanino.

4. En un test distribuído normalmente, el sujeto A ocupa el centil 20, el B el centil 40 y el C el centil 60. Por lo tanto, la diferencia entre las puntuaciones directas de A y B será la misma que para los sujetos B y C. V ( ) F ( ) Depende ( ). Razone su respuesta:

5. Las puntuaciones de una persona en tres escalas diferentes han sido: 60, 70 y 80. Diga razonadamente qué puntuación corresponde a cada escala:

a) La puntuación en la escala centil es _____
b) La puntuación en la escala T es _____
c) La puntuación en la escala D es _____

6. En un grupo normativo se han obtenido los estaninos (normalizados) y las puntuaciones típicas normalizadas de cada persona. Entre ambos se obtiene una correlación de 1. Diga cuál de las siguientes alternativas es correcta y porqué.

a) No es posible esa correlación.
b) Se ha obtenido por casualidad.
c) Es necesariamente 1.
d) Sólo es 1 si la distribución es simétrica.

7. La puntuación de una persona en un test de inteligencia se encuentra 0,5 desviaciones típicas por encima de la media del grupo normativo. Obtenga sus puntuaciones en los baremos Z, T y D.

8. Algunas de las puntuaciones de Juan y Antonio en un examen han sido las siguientes:

	Centil	Z_n	E_n	D_n	T_n
Juan	93	1,5			
Antonio					65

Complete las puntuaciones omitidas.

9. En un test cuyas puntuaciones se distribuyen normalmente, 5 personas (numeradas del 1 al 5) obtienen las siguientes puntuaciones en los correspondientes baremos:

1) $D_n = 50$ 2) $T_n = 20$ 3) $E_n = 5$ 4) $Z_n = -3$ 5) Centil = 90

Sitúe el número correspondiente a cada persona en la curva normal.

10. Aplicamos un test a un grupo normativo de 350 personas. La distribución de frecuencias resultante fue:

X	45	44	43	42	41	40	39	38
f	5	15	45	85	90	56	44	10

a) ¿Qué centil, puntuación típica y típica normalizada corresponden a la persona que obtenga una puntuación directa de 42?
b) Sabemos que la persona A en la escala T tiene una puntuación que coincide con la de la persona B en la escala D. ¿Han podido tener las dos personas la misma puntuación en el test?

11. Dos personas tienen exactamente los mismos conocimientos en un examen de "Introducción a la Psicometría", que consta de 50 preguntas con 4 alternativas de respuesta y sólo una correcta. La primera, una persona poco amante del riesgo, sólo responde a lo que sabe, y obtiene 30 aciertos. La segunda, mucho más arriesgada, responde a todas las preguntas.

a) ¿Cúantos fallos es previsible que tenga la segunda persona si responde completamente al azar las preguntas que no sabe? ¿Cuál será su puntuación después de descontarle los aciertos aleatorios?.
b) Obtenga e interprete el estanino en el que se encontrará la primera persona si su puntuación típica normalizada en el examen fue de –1.5.

12. Una persona obtiene una puntuación T_n = 80 en el baremo de una escala de autoritarismo realizado en una muestra A. La misma persona obtiene una T_n = 70 en la misma escala baremada en una muestra B. ¿Cuál de las dos muestras manifiesta mayor nivel de autoritarismo? Razone su respuesta.

SOLUCIONES

1. La construcción de baremos tiene por objeto poder interpretar puntuaciones directas de los tests en función de la posición relativa que esas puntuaciones directas tienen en el conjunto de las puntuaciones obtenidas.

2. a) Los centiles se muestran en la cuarta fila

X	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
f	2	3	4	11	17	15	23	20	19	23	22	15	8	8	4	5	1
F	2	5	9	20	37	52	75	95	114	137	159	174	182	190	194	199	200
C	1	3	5	10	19	26	38	48	57	69	80	87	91	95	97	100	100

b)

X	Z_x	T	D
10	-1,43	35,7	21,4
21	1,86	68,6	87,2

c)

X	Z_n	T_n	D_n	E
10	-1,28	37,2	24,4	2,44
21	1,88	68,8	87,6	8,76

3. a) C_{21}= 40
b) Z_n = -0,8
c) D = 40
d) E = 3

4. Falso. La escala de centiles tiene propiedades ordinales. Si, como se dice, la distribución es normal, la diferencia de 20 en la escala de centiles extremos indicará una mayor diferencia de puntuaciones que la diferencia de 20 en centiles centrales. La diferencia entre A y B será mayor que la diferencia entre B y C.

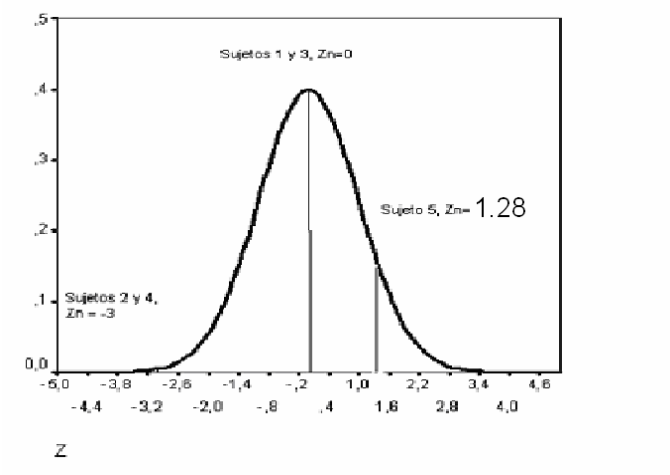
5. T = 60 D = 70 C_{80}

6. La alternativa correcta es la c), ya que ambos baremos resultan de una transformación lineal de las puntuaciones típicas normalizadas.

7. Z = 0.5, T = 50 + (10) 0.5 = 55 y D = 50 + (20) 0.5 = 60.

8. Puntuaciones de Juan: E_n = 8, D_n = 80, T_n = 65.
Puntuaciones de Antonio: E_n = 8, D_n = 80, Z_n = 1.5 y centil 93.

9.



10.

X	45	44	43	42	41	40	39	38
f	5	15	45	85	90	56	44	10
F	350	345	330	285	200	110	54	10
C	100	98,57	94,28	81,43	57,14	31,43	15,42	2,85

a) $C_{81} = 42$, $Z = 0{,}54$, $Z_n = 0{,}89$
b) Sólo si $Z_A = Z_B = 0$. En ese caso, $T_A = D_B = 50$

11. a) La segunda persona tendrá 15 errores y 5 aciertos (15 = (20)3/4 y 5 = (20)/4). Su puntuación corregida será 30.
b) El estanino será 2.

12. La persona tiene una puntuación mayor en la muestra A que en la muestra B, eso indica que la primera muestra es menos autoritaria que la muestra B.

TEMAVI: INTRODUCCIÓN A LA TEORÍA DE LA RESPUESTA AL ÍTEM

1.- INTRODUCCION

La Teoría de la Respuesta al Item (TRI) constituye un nuevo enfoque en Psicometría que permite superar algunas de las limitaciones de la Teoría Clásica de los tests (TC).

Su propósito es similar al de la Teoría clásica. Pretende obtener la puntuación que corresponde a una persona en una dimensión o rasgo, como su inteligencia, su nivel en un cierto rasgo de personalidad, su dominio en una cierta materia, etc..

La TRI debe su nombre a que se centra más en las propiedades de los ítems individuales que en las propiedades globales del test, como hacía la TC.

Este capítulo es sólo una breve introducción a la TRI. Lo que vamos a estudiar sólo es aplicable a ítems que puedan cuantificarse como cero o uno. La TRI permite también el análisis de ítems con otros formatos de respuesta (por ejemplo, las categorías ordenadas), pero tales desarrollos no son tratados en estas líneas.

Buena parte de la Psicometría actual está relacionada con la TRI y es muy abundante la bibliografía existente. Las personas interesadas en ampliar conocimientos, pueden consultar los libros de Muñiz (1997), Hambleton, Swaminathan y Rogers (1991) y Hambleton y Swaminathan (1985). Nos hemos basado principalmente en los dos primeros para redactar este tema.

Entre las principales limitaciones de la TC se pueden exponer las siguientes:

a) La principal limitación consiste en que las características del test y las puntuaciones de las personas no pueden ser separados: Se define la puntuación de una persona como el número de preguntas que acierta, y la dificultad de un ítem como la proporción de personas que lo responden correctamente en un determinado grupo. Esto tiene una serie de consecuencias negativas:

- Las características de los ítems dependen del grupo de personas en el que se han aplicado. Por ejemplo, supongamos que queremos conocer el índice de dificultad de un determinado ítem que mida conocimientos de tauromaquia. Dicho índice será muy diferente si utilizamos en la baremación un grupo de personas abonadas a la feria de San Isidro o un grupo de turistas japoneses.

- La puntuación de una persona depende del conjunto particular de ítems administrados. La puntuación que una persona obtenga será diferente si le aplicamos dos tests que midan la misma característica pero cuyo nivel de dificultad sea diferente. Esto hace muy difícil comparar dichas puntuaciones, que sólo podrán interpretarse en relación al test en el que fueron obtenidas.

Frente a la TC, una de las propiedades de la TRI es su **invarianza**, en un doble sentido: invarianza de los ítems respecto a posibles diferentes distribuciones de la habilidad o del rasgo (en lo sucesivo nivel de habilidad y de rasgo serán sinónimos), e invarianza de la habilidad medida a partir de diferentes conjuntos de ítems. Haremos un breve comentario sobre cada tipo de invarianza. Si las condiciones de aplicación de la TRI se cumplen, ha de ocurrir lo siguiente:

- Sea cual sea la distribución de los niveles de rasgo obtendremos las mismas estimaciones de los parámetros de los ítems. Esta propiedad se cumple también en otros ámbitos. Por ejemplo, en Estadística, si se cumplen los supuestos de la regresión lineal, se llega a los mismos parámetros cuando se ajusta la recta de regresión a toda la población o sólo a parte de ella. Análogamente, los parámetros de los ítems deberán ser los mismos si éstos se han aplicado a un grupo de personas con alto nivel de rasgo, o a un grupo con niveles bajos. Es decir, los parámetros de los ítems serán los mismos sea cual sea la distribución de los niveles de habilidad de la muestra en los que se han aplicado.

- El nivel de habilidad de una persona puede ser obtenido a partir de conjuntos de ítems distintos. Algunas de las aplicaciones de la TRI descansan precisamente en esta propiedad (véase más adelante el apartado "Aplicaciones").

b) Una segunda limitación tiene con ver el error de medida. La TC supone que el error de medida es una propiedad del test y, por lo tanto, igual para todos los sujetos, independientemente de cual sea su puntuación. Por el contrario, la TRI permite obtener la precisión con la que cada persona es medida.

La TRI permite superar estas y otras limitaciones de la TC mediante unos supuestos fuertes y restrictivos, y una metodología más compleja, que requiere establecer modelos matemáticos, la estimación de sus parámetros, enjuiciar el ajuste entre datos y modelos, etc..

Antes de ver cuales son los supuestos, vamos a estudiar los principales modelos de la TRI.

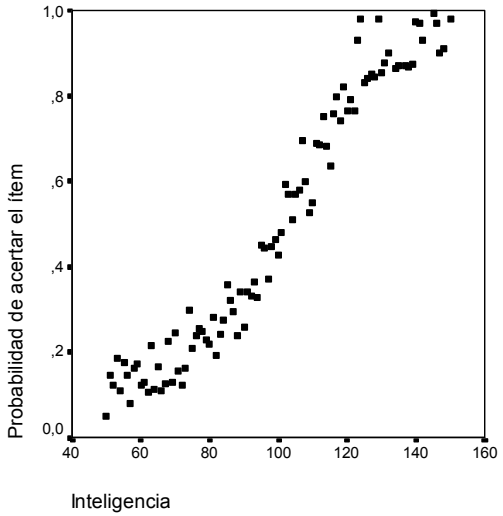
2.- CURVA CARACTERÍSTICA DEL ÍTEM

La curva característica de un ítem (CCI) indica la probabilidad que tienen de acertarlo las personas que se enfrentan a él. Esta probabilidad depende, naturalmente, de cual sea el nivel de la persona en la variable medida.

Podemos ver esto con más claridad mediante un ejemplo. Supongamos que tenemos un test que mide inteligencia y que ha sido aplicado a muchísimas personas (100.000, por ejemplo). Supongamos que la menor y mayor puntuación obtenidas en el test son 50 y 150. Vamos a representar el rendimiento en un ítem concreto de la siguiente forma: Nos fijamos en todas las personas que han obtenido la puntuación 50 (supongamos que son 132). Vemos cuantas personas de las anteriores han acertado el ítem (supongamos que han sido sólo 5) y

calculamos la proporción (5/132 = 0.04). Hacemos lo mismo con los que obtuvieron en el test 51 puntos (y obtenemos la proporción, supongamos que 0.15),... con las que obtuvieron en el test 100 (la proporción fue 0.45),... con las que obtuvieron 150 (la proporción fue 0.99). La siguiente gráfica muestra la proporción de aciertos en el grupo de personas que obtuvo en el test 50 puntos, 51, .. 150.

Gráfica 1



En este ejemplo podemos ver que cuanto mayor es el cociente intelectual de las personas, mayor es la proporción de aciertos en el ítem. A una puntuación de 100 le corresponde una proporción de 0.45; mientras que a una de 150 le corresponde una proporción de 0.99.

En la gráfica 1 tenemos una CCI empírica, pero la TRI necesita resumir la información que contiene cada CCI empírica en una fórmula o modelo en el que uno, dos o tres valores resuman la información contenida en la CCI empírica. En la aplicación de de la TRI, un paso inexcusable es optar por un modelo (o fórmula) que sea una buena descripción del rendimiento en los ítems. Vamos a ver a continuación que son varios los modelos que podrían dar cuenta de una distribución como la mostrada en la gráfica 1. Los modelos de CCI más utilizados en la práctica son los logísticos de uno, dos y tres parámetros.

2.1. – MODELO LOGÍSTICO DE UN PARAMETRO (MODELO DE RASCH)

Este es el modelo más simple de todos. Se le llama también modelo de Rasch. La probabilidad de acertar un ítem depende solamente del nivel de dificultad de dicho ítem y del nivel del sujeto en la variable medida (nivel de rasgo o habilidad).

La expresión matemática es:

$$P(\theta) = \frac{e^{D(\theta - b)}}{1 + e^{D(\theta - b)}} = \frac{1}{1 + e^{-D(\theta - b)}}$$

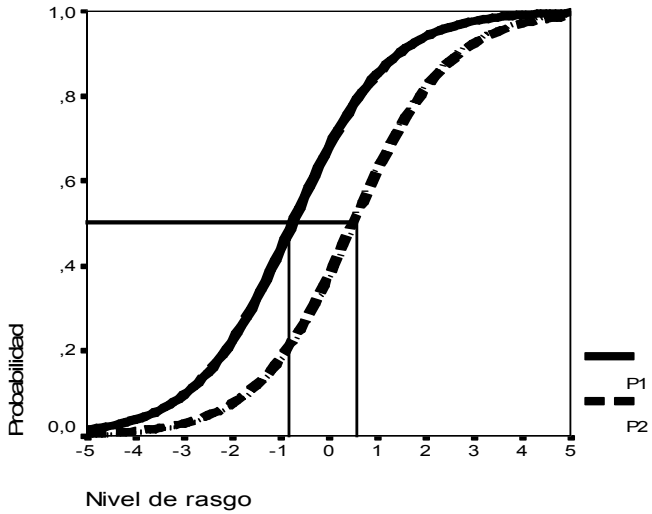
Donde

P(θ): Probabilidad de acertar el ítem si el nivel de rasgo es θ.
θ : Nivel de habilidad del sujeto.
b : Indice de dificultad del ítem.
e : Base de los logaritmos neperianos (2.718)
D : Constante (D = 1.7 ó 1)

El nivel de habilidad del sujeto (θ) puede definirse en cualquier escala (en la gráfica 1 se ha utilizado la escala de cociente intelectual). No obstante, en la práctica, suele utilizarse una escala típica, con media cero, varianza uno y un rango de valores entre -3.0 y 3.0.

El índice de dificultad (b) es aquel valor de θ para el cual P(θ)= 0.5. Por tanto, cuanto mayor sea "b" más difícil es el ítem. En la gráfica 2, se han representado dos CCIs. En la primera, la que está más a la izquierda, el valor de θ al que corresponde P(θ)= 0.5 es aproximadamente -0.95. Por lo tanto, la dificultad del primer item es b_1= -0.95. En el segundo ítem, el valor de θ al que corresponde P(θ)= 0.5 es aproximadamente 0.6. Por lo tanto, la dificultad del segundo ítem es b_2= 0.6. La gráfica muestra que la probabilidad de acertar el ítem es sistemáticamente menor en el ítem 2 que en el ítem 1. El ítem 2 es más difícil que el uno, y sus índices de dificultad así lo muestran ($b_2 > b_1$).

Gráfica 2



2.1.- MODELO LOGÍSTICO DE DOS PARAMETROS

Este modelo añade al anterior un segundo parámetro que indica la capacidad discriminativa del ítem:

$$P(\theta) = \frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}} = \frac{1}{1 + e^{-Da(\theta - b)}}$$

Donde a" es el índice de discriminación del ítem.

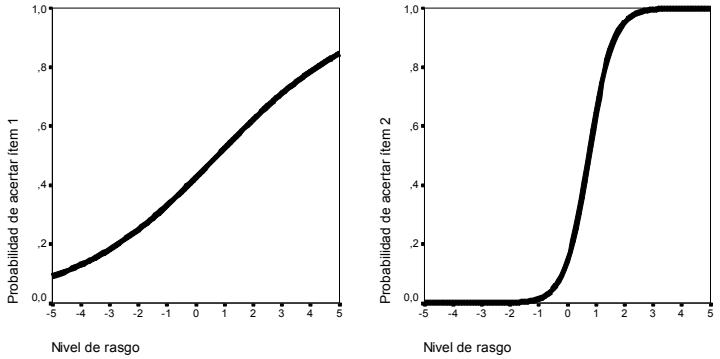
El parámetro "a" indica la mayor o menor inclinación o pendiente de la CCI cuando θ=b. Normalmente los valores de "a" oscilan entre 0,3 y 2.5, y se suelen considerar items "discriminativos" los que tienen valores "a" mayores de uno.

En la gráfica 3 vemos la CCI de dos ítems de igual dificultad (b_1= b_2= 0.75), la principal diferencia entre ellos es que el ítem 2 (el de la derecha), cuando θ = 0.75, tiene una pendiente mucho mayor (a_2 = 2.4) que la del ítem 1 (a_1 = 0.4). Como la pendiente es tan alta, las personas con θ > 0.75 tienen casi todas ellas una muy alta probabilidad de acertar el ítem 2 (y casi todas ellas lo acertarán), y las personas con θ < 0.75 tienen casi todas ellas una

probabilidad próxima a cero de acertarlo (y casi ninguna lo acertará). Por lo tanto, el ítem 2 **discrimina** entre los que tienen $\theta > 0.75$ y los que tienen $\theta < 0.75$.

Por su parte, el ítem 1 tiene muy poca pendiente cuando $\theta = 0.75$. En consecuencia, aunque la mayoría de las personas con $\theta > 0.75$ lo acertarán, muchas lo fallarán (pues la probabilidad de acierto es claramente inferior a uno). Igualmente, aunque la mayoría de las personas con $\theta < 0.75$ fallarán el ítem, muchas lo acertarán, pues la probabilidad de acierto es claramente superior a cero. En el ítem 1 la probabilidad crece muy suavemente a medida que aumenta θ por lo que no es buen discriminador entre las personas con $\theta > 0.75$ y las que tienen $\theta < 0.75$.

Gráfica 3



2.3.- MODELO LOGÍSTICO DE TRES PARÁMETROS

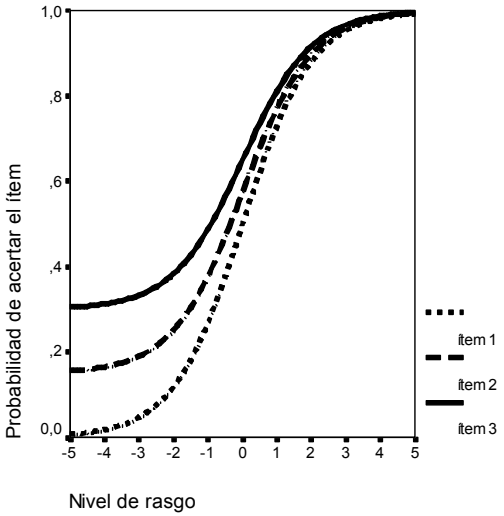
Este modelo añade a los dos parámetros "a" y "b" un tercero, "c", que representa la probabilidad de acertar el ítem al azar. Exactamente "c" es el valor de $P(\theta)$ para valores extremadamente bajos de θ. La expresión matemática es la siguiente:

$$P(\theta) = c + \frac{(1 - c)\, e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}}$$

$$= c + \frac{1 - c}{1 + e^{-Da(\theta - b)}}$$

En la gráfica 4 podemos ver la CCI de varios ítems con los mismos valores de a (1) y b (0), pero distintos valores de parámetro "c" ($c_1 = 0$, $c_2 = 0.15$ y $c_3 = 0.30$).

Gráfica 4



3.- SUPUESTOS DE LA TRI

3.1.- UNIDIMENSIONALIDAD

Tal y como hemos visto en el apartado anterior, en todos los modelos de CCIs, la probabilidad de acertar un ítem depende únicamente de sus parámetros y de θ. En un ítem que mida el nivel de vocabulario de inglés, la probabilidad de acertarlo depende de los valores "a", "b" y "c" del ítem y del nivel de vocabulario inglés de la persona (su θ). La CCI excluye que el rendimiento en el ítem dependa de los niveles de la persona en otros rasgos más o menos relacionados con el de vocabulario de inglés (como el nivel de inglés hablado, el nivel de gramática inglesa, ...). De tener en consideración otros rasgos, en la fórmula aparecerían los elementos θ_2, θ_3, etc.. , es decir, los niveles de la persona en esos otros rasgos. En otras palabras, el rendimiento en un ítem depende del nivel de la persona en un **sólo** rasgo o dimensión.

Un test consta de un conjunto de ítems. La TRI supone además que **todos** los ítems que forman el test han de medir **un mismo y único rasgo**. El supuesto de unidimensionalidad exige que todos y cada uno de los ítems del test midan una única dimensión.

Este supuesto nunca se cumple totalmente porque el rendimiento en un test está influido por variables cognitivas y de personalidad, como la motivación, ansiedad, etc. Por lo que, en la práctica, es una cuestión de grado, y no puede afirmarse categóricamente si un conjunto de ítems son o no unidimensionales. Hay, no obstante, varios métodos para comprobar la unidimensionalidad. El propuesto por Reckase (1979) se basa en el análisis factorial y consiste en estudiar la varianza explicada por el primer factor extraído de la matriz de correlaciones entre ítems. En la práctica, cuando el primer factor explica más de un 25% de la varianza total, tras haber eliminado los ítems con saturaciones inferiores a 0.10, se considera que se cumple el supuesto de unidimensionalidad.

3.2.- INDEPENDENCIA LOCAL

Existe independencia local entre los ítems de un test si la respuesta que una persona da a uno de ellos no depende de las respuestas que da a los otros.

La independencia local se deriva de la unidimensionalidad porque, simplemente, significa que la respuesta a un ítem sólo depende de sus parámetros y de θ, y no está influida por el orden de presentación de los ítems, las respuestas que ya se hayan dado, etc..

Matemáticamente puede expresarse diciendo que la probabilidad de que un sujeto acierte "n" ítems es igual al producto de las probabilidades de que acierte cada uno de ellos por separado.

Por ejemplo, un test consta de dos ítems y la probabilidad de que Juan acierte el primero es $P_1 = 0.4$ y la de que acierte el segundo $P_2 = 0.8$. El principio de independencia local establece que la probabilidad de que acierte los dos viene dada por: $(P_1)(P_2) = (0.4)(0.8) = 0.32$.

La probabilidad de acertar el primero y fallar el segundo sería (como $Q_2 = 1 - P_2 = 1 - 0.8 = 0.2$):

$$(P_1)(Q_2) = (0.4)(0.2) = 0.08.$$

La de que falle el primero y acierte el segundo será $(Q_1)(P_2) = (0.6)(0.8) = 0.48$.

La de que falle ambos ítems será $(Q_1)(Q_2) = (0.6)(0.2) = 0.12$.

Supongamos que 100 personas con idéntico nivel de rasgo que Juan contestan al test. Espereramos aproximadamente los siguientes resultados (1, acierto; 0, error):

Ítem 1	Ítem 2	Número de personas
1	1	32
1	0	8
0	1	48
0	0	12
		100

Si correlacionamos las cien respuestas al primer item con las cien respuestas al segundo, la correlación de Pearson es cero. Lo visto sugiere un procedimiento para contrastar si el supuesto de independencia local se cumple. Consiste en obtener la matriz de correlaciones entre los ítems, pero no en la muestra completa, sino en submuestras que sean lo más homogéneas posible en cuanto al nivel de habilidad de sus miembros. En tales submuestras tiene que ocurrir que ningún ítem correlacione con ningún otro, si se cumple el supuesto. (Hambleton y otros, 1991, pag. 56).

4.- ESTIMACIÓN DE PARÁMETROS

Seleccionado un modelo de TRI, hay que aplicar el test a una muestra amplia y estimar los parámetros de cada ítem y la θ de cada sujeto, a partir de la matriz de respuestas obtenidas. Si tenemos, por ejemplo, diez ítems que miden un mismo rasgo, los podemos aplicar a una muestra de 300 personas. La matriz de datos tendrá 300 filas, siendo cada fila la secuencia de unos (aciertos) y ceros (errores) de cada persona de la muestra. Si queremos aplicar el modelo logístico de tres parámetros, tendremos que estimar los 30 parámetros de los ítems (es decir, "a", "b" y "c" de cada ítem) y 300 parámetros de las personas (los 300 valores de "θ", uno por persona). La estimación de parámetros es el paso que nos permite llegar de las respuestas conocidas de las personas a los ítems a los valores desconocidos de los parámetros de los ítems y de los niveles de rasgo.

Para obtener las estimaciones se aplica fundamentalmente el **método de máxima verosimilitud**. La lógica general de la estimación consiste en encontrar los valores de los parámetros que hagan más probable la matriz de respuestas obtenida.

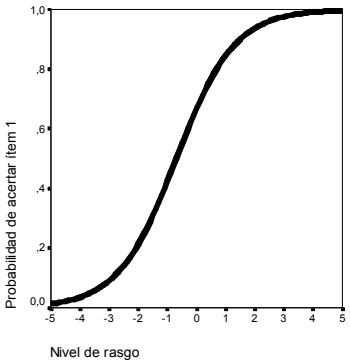
Si lanzamos una moneda diez veces y obtenemos siete caras, el estimador máximo-verosímil del parámetro "p" (probabilidad de cara de la moneda) es 7/10= 0.7, como se demuestra en los libros de Estadística (véase Amón (1984), pag. 249 y ss). El resultado "siete caras en diez lanzamientos" es poco compatible con que la probabilidad de cara sea 0.1, ó 0.2, ... De hecho, la probabilidad de obtener siete caras y tres cruces es prácticamente cero si p= 0.1 o si p= 0.2. Dicha probabilidad pasa a ser 0.117 si p= 0.5, y alcanza el máximo valor (0.267) cuando p= 0.7. El estimador máximo-verosímil proporciona el valor de "p" bajo el que tiene máxima probabilidad el suceso que hemos encontrado.

En TRI, el procedimiento de estimación sigue una lógica similar. Se obtienen las estimaciones de los parámetros y de los niveles de θ con los que la matriz de datos encontrada tiene la máxima compatibilidad.

Supongamos, por ejemplo, que tenemos un test compuesto por tan sólo dos ítems, y se lo aplicamos a un sujeto. Supongamos también que acierta el primero y falla el segundo. A partir de estas respuestas, la estimación máximo-verosímil de su θ se puede explicar de forma gráfica, como lo hacemos a continuación (en este ejemplo, para simplificar la explicación, suponemos que los parámetros de los ítems son conocidos).

Como el sujeto ha acertado el primer ítem, podemos calcular, mediante su CCI (recuérdese que los parámetros del ítem son conocidos), la probabilidad de que esto ocurra para cada nivel de θ. Gráficamente, para un ítem cuyo único parámetro es $b_1 = -0.7$:

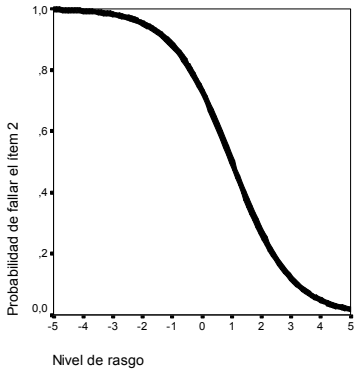
Gráfica 5



Si sólo hubiera respondido a ese ítem, a partir de la gráfica anterior podemos ver que no existe un único valor de θ para el que la probabilidad del suceso encontrado (acierto en el primer ítem) sea máxima. Por el contrario, son infinitos los valores de θ que para los que la CCI alcanza el valor máximo 1.

Como el sujeto ha fallado el segundo ítem, a partir de su CCI podemos calcular la probabilidad de que esto ocurra para cada uno de los valores de θ. En concreto, como la probabilidad de fallar (Q) se puede obtener a partir de la probabilidad de acertar (Q = 1-P), podremos representar la probabilidad de error en el segundo ítem como se muestra en la siguiente gráfica. Nótese que la siguiente gráfica no es la CCI del ítem 2, pues para cada valor de θ se ha representado la probabilidad de **error** y no la de acierto, como exige la CCI. Supongamos que el único parámetro del ítem 2 es $b_2 = 1$.

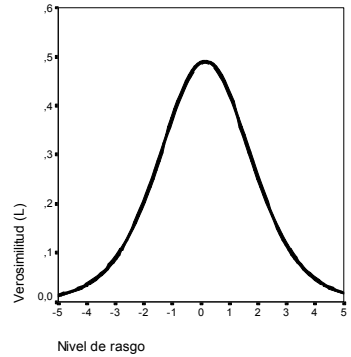
Gráfica 6



Esta gráfica nos indica que es más probable que fallen el ítem los sujetos con niveles bajos de habilidad que los sujetos con niveles altos (cosa bastante lógica). Por lo tanto, si el sujeto sólo hubiese respondido a este ítem, de nuevo son infinitos los valores de θ que maximizan la probabilidad del suceso encontrado (error en el segundo ítem).

Como hecho ha respondido a dos ítems, el valor estimado de θ para este sujeto sería aquel que haga más probable el resultado obtenido (acertar el primer ítem y fallar el segundo). Según el supuesto de independencia local, ambos sucesos son independientes y, por lo tanto, la probabilidad de que ocurran ambos conjuntamente es igual al producto de las probabilidades de acertar el primero (P_1) por la de fallar el segundo (Q_2). Si representamos gráficamente la función $L = (P_1)(Q_2)$ para cada valor de θ, correspondiente al ejemplo que venimos comentando, obtendríamos una curva parecida a la siguiente:

Gráfica 7



En este caso vemos que la θ que hace más probable el resultado obtenido (acierto en el primer ítem y fallo en el segundo) es algo mayor que cero. De hecho, 0.15 será la θ estimada para este sujeto.

En general, una persona responderá a un número de ítems mayor de dos y producirá una particular secuencia de unos y ceros. La probabilidad de obtener tal secuencia de aciertos y errores se puede escribir como:

$$L = \prod P^R Q^{1-R}$$

Donde:

R: Resultado en cada ítem (1, acierto; 0, fallo)
P: Probabilidad de acierto en cada ítem
Q: Probabilidad de error en cada ítem (Q= 1-P).

La θ estimada por el método de máxima verosimilitud será el valor de θ para el que la anterior expresión alcanza su máximo valor.

Apliquemos lo anterior al siguiente **ejemplo**. Un test consta de 4 ítems, cuyos parámetros, según el modelo de Rasch, son -1, 0, 1 y 2. Una persona completa el test y acierta los tres primeros ítems y falla el cuarto. Obtenga el valor de la función de verosimilitud, L, para los siguientes valores de θ: -3, -2, -1, 0, 1 y 2. ¿Cuál de los anteriores valores maximiza L?

Aplicando la fórmula del modelo de Rasch (o de un parámetro), se obtiene la probabilidad de acierto para cada ítem y cada uno de los valores de θ:

		P(θ)						
Ítems	b	-3	-2	-1	0	1	2	3
1	-1	0.03	0.15	0.50	0.85	0.97	0.99	0.99
2	0	0.01	0.03	0.15	0.50	0.85	0.97	0.99
3	1	0.01	0.01	0.03	0.15	0.50	0.85	0.97
4	2	0.01	0.01	0.01	0.03	0.15	0.50	0.85

La función de verosimilitud, L, al haber acierto en los 3 primeros ítems y fallo en el último, será la siguiente:

$$L = (P_1^1 Q_1^0)(P_2^1 Q_2^0)(P_3^1 Q_3^0)(P_4^0 Q_4^1) = (P_1)(P_2)(P_3)(Q_4)$$

Aplicando la fórmula anterior a cada uno de los valores de θ se obtienen los siguientes resultados:

L(3) = (0.99)(0.99)(0.97)(1-0.85) = 0.14
L(2) = (0.99)(0.97)(0.85)(1-0.50) = 0.41
.
.
.

Los restantes valores de L son L(1)= 0.35, L(0)= 0.06, L(-1) = L(-2) = L(-3) = 0.0. Por lo tanto, de los siete valores de θ considerados, el valor que maximiza L es θ = 2.

Cuando se trata de estimar en una situación real el nivel de rasgo, no se hace una búsqueda restringida a unos cuantos valores, se necesita hallar el valor de θ que maximiza L de entre todos los posibles valores, no sólo de entre unos pocos.

En el caso de la TRI no existen fórmulas que permitan obtener las estimaciones de manera directa. En el ejemplo de las monedas se sabe que el estimador máximo-verosímil de la proporción poblacional es la proporción muestral. En la TRI, al no existir tales fórmulas, las estimaciones se obtienen por métodos numéricos, mediante programas de ordenador. En el caso más general se establece una función L que depende de los parámetros de los ítems y de los niveles de rasgo. Los programas de ordenador contienen algoritmos que encuentran el conjunto de estimaciones para el que la función L alcanza el valor máximo. Los parámetros de los ítems y los niveles de rasgo de las personas serán los valores dados por el programa de ordenador para una matriz de respuestas particular.

En la Teoría Clásica, una vez aplicados unos ítems a un conjunto de personas, se puede obtener la puntuación de cada persona en el test combinando las puntuaciones en los ítems del test. En la TRI, una vez que se han aplicado los ítems, se genera la matriz de respuestas que contiene los aciertos y fallos de cada persona en cada ítem del test. A continuación, se ha de aplicar un programa de ordenador (ASCAL, BILOG,..) que nos dará los niveles de rasgo y los parámetros de los ítems. Según hemos visto, por tratarse de estimaciones por el método de máxima verosimilitud, los valores que nos da el programa son los que hacen más plausible la matriz de datos original, son los más compatibles con la matriz de datos original.

5.- FUNCIÓN DE INFORMACIÓN

Una vez aplicado un conjunto de ítems y estimado el nivel de habilidad de un sujeto, la TRI nos permite calcular el **error típico de estimación** (Se) de esa persona en el test aplicado. Esto es una diferencia fundamental con la TC, que asume que el error es el mismo para todos los sujetos.

El error típico de estimación nos dice la precisión con que hemos estimado θ. A mayor error, menos precisión. Su tamaño depende de varios factores:

1- Número de ítems aplicado: En general, al aumentar la longitud del test disminuye Se.

2- La capacidad discriminativa de los ítems: Al aumentar el parámetro "a" disminuye Se.

3- La diferencia entre "b" y θ: Cuanto más próximo a θ esté el índice de dificultad de los ítems (b), menor será Se.

La varianza de las puntuaciones θ estimadas, Var(θ), se obtiene mediante la expresión siguiente:

$$Var(\theta) = S_e^2 = \frac{1}{\Sigma \frac{(P')^2}{PQ}}$$

Donde P' es la derivada de P. La varianza anterior nos dice cómo es de importante la variación entre los valores de θ estimados y el valor verdadero de θ. Cuanto menor sea esta varianza, indicará que más nos podemos fiar del test; pues sabemos que son pocas las diferencias entre los valores estimados y el verdadero.

Por su parte, el error típico de estimación de θ es la desviación típica de las puntuaciones θ estimadas, es decir,

$$S_e = \sqrt{S_e^2}$$

El error típico de estimación permite obtener el intervalo confidencial en el que, con probabilidad predeterminada, se ha de encontrar el nivel de habilidad de la persona. En concreto, si a la "θ" estimada de una persona le sumamos y restamos (1.96)Se, obtenemos los extremos del intervalo en el que, con probabilidad 0.95, se encontrará su verdadero nivel de rasgo.

Por ejemplo, si la θ estimada es 0.8 y su error típico de estimación es 0.22, entonces, el nivel de rasgo de dicha persona se encuentra entre 0.37 (pues 0.8-(1.96)(0.22)= 0.37) y 1.23 (pues 0.8+(1.96)(0.22) = 1.23), con probabilidad 0.95.

La **función de información** del test aplicado se define como la inversa de Var(θ), es decir:

$$I(\theta) = \frac{1}{S_e^2}$$

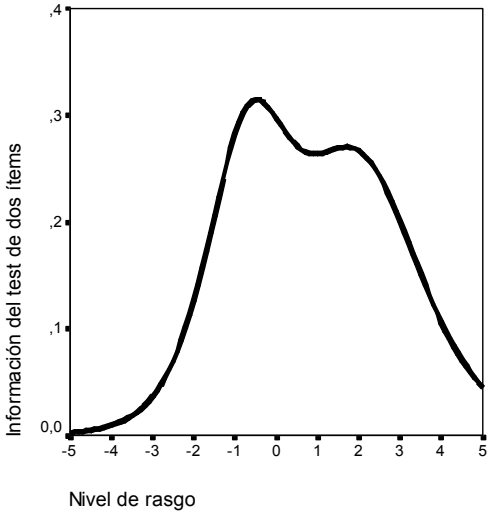
Por lo tanto,

$$I(\theta) = \Sigma \frac{(P')^2}{PQ}$$

Cuanto mayor sea I(θ) menor será Se y, por tanto, mayor la precisión de la estimación de θ.

Si se calcula I(θ) para todos los niveles de θ y se representa gráficamente se obtiene una curva como la que muestra la siguiente gráfica:

Gráfica 8



Vemos que este test (compuesto por dos ítems, cuyos parámetros son a_1= 1.5 , b_1 = -0.7 a_2= 1 y b_2 = 2) aporta más información para valores de θ en torno a -0.5.

La FI tiene una gran importancia en la utilización de los tests, ya que nos permite elegir aquel que aporte más información en el intervalo de θ que estemos interesados en medir.

También es muy útil en la construcción del test. A partir de un banco de ítems calibrados (es decir, de los que hemos estimado sus parámetros) podemos seleccionar aquellos que permitan que la FI se ajuste a unos objetivos determinados.

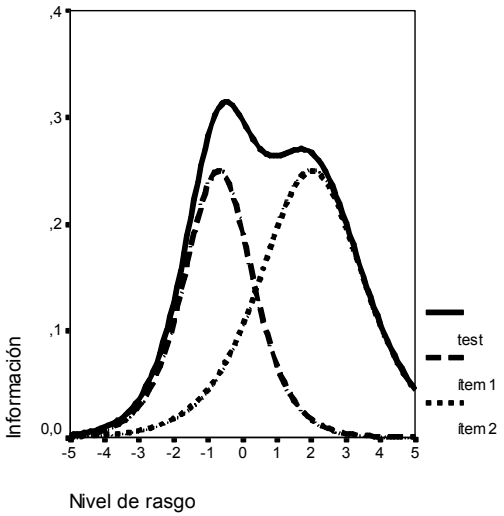
Todos los conceptos anteriores referidos a la función de información del test son aplicables también a cada uno de los ítems por separado. De hecho la FI del test no es más que la suma de las FFII de cada uno de los ítems que lo componen. En concreto la FI de un ítem sería:

$$I(\theta) = \frac{(P')^2}{PQ}$$

Vemos que la única diferencia con la FI del test es que ha desaparecido el signo de sumatorio.

Al igual que con el test completo, podemos representar gráficamente la FI de los ítems y ver a que nivel de θ proporcionan más información. La siguiente gráfica muestra la función de información de los dos ítems que forman el test y la función de información del test.

Gráfica 9



Esto nos permite elegir los ítems más adecuados en cada momento en función de nuestras necesidades. Por ejemplo, si queremos llevar a cabo una selección de personal en la que sólo vamos a elegir unos pocos sujetos muy competentes, a partir de un banco de ítems previamente calibrado, podríamos elegir aquellos ítems que proporcionan más información para niveles altos de θ. Esto nos permite reducir enormemente el número de ítems de un test sin perder precisión al estimar θ.

6.- APLICACIONES

La TRI ha permitido la elaboración y el desarrollo de tests adaptativos informatizados (TAIs) (véase Renom, 1993; Olea, Ponsoda y Prieto, 1997; Olea y Ponsoda, 2003). Tales tests

difieren sustancialmente de los tests al uso. Un TAI consta de un banco de ítems bien calibrado y de un programa de ordenador encargado de decidir qué ítem del banco presentar a la persona, de presentárselo, de analizar la respuesta emitida por la persona, de elegir un nuevo ítem del banco, etc..

Un TAI difiere muchísimo de un test de lápiz y papel. Una primera diferencia es que es administrado por un ordenador y una segunda es que cada persona es evaluada con items distintos. Sin embargo, lo fundamental de los TAIs es que los ítems son elegidos con el criterio de estimar el nivel de habilidad de la persona con la máxima precisión y menor número de ítems. Más en concreto, un TAI procede como se expone a continuación:

a) Presentación del primer ítem.
b) Estimación del nivel de rasgo de la persona.
c) Búsqueda del ítem del banco más informativo para el nivel de θ estimado en el paso precedente.
d) Aplicación del ítem elegido.
e) Estimación del nivel de rasgo correspondiente a la secuencia de respuestas dada a los ítems presentados.
f) De nuevo paso "c", y así sucesivamente hasta que se haya conseguido un error típico de estimación menor que un tope preestablecido o se haya administrado un predeterminado número de ítems.

El principal logro de los TAIs es que con muy pocos ítems (veinte, más o menos) se pueden conseguir precisiones en la medición comparables o mejores que las obtenidas en tests no adaptativos mucho más largos. Esto es así porque en los TAIs sólo se administran ítems auténticamente informativos para determinar el nivel de rasgo de la persona y se evitan los ítems demasiados fáciles o difíciles, que apenas informan sobre el nivel de rasgo. Hemos construido un TAI de vocabulario inglés (Ponsoda, Olea y Revuelta, 1994) y hemos obtenido que, en ocasiones, con sólo diez ítems se obtiene una excelente precisión (un error típico de estimación equivalente a un coeficiente de fiabilidad de 0.9).

7.- REFERENCIAS (de este tema)

Amón J. (1984). *Estadística para psicólogos. Probabilidad. Estadística Inferencial*. Volumen 2. 3ª edición. Madrid: Pirámide.

Hambleton R.K. y Swaminathan H. (1985). *Item Response Theory: Principles and applications*. Boston: Kluwer.

Hambleton R.K, Swaminathan H. y H.J. Rogers (1991). *Fundamentals of Item Response Theory*. MMSS volumen 2. Londres: Sage.

López Pina, José Antonio (1995). *Teoría de la respuesta al ítem: fundamentos*. Barcelona: PPU. Barcelona.

Muñiz Fernández J. (1997). *Introducción a la Teoría de Respuesta a los Items*. Madrid: Pirámide.

Olea, J., Ponsoda, V. y Prieto, G. (1997). *Tests informatizados*. Madrid: Pirámide.

Olea, J. y Ponsoda, V. (2003). *Tests adaptativos informatizados*. Madrid: UNED Ediciones.

Ponsoda V., Olea J. y Revuelta J. (1994). ADTEST: A computer adaptive test based on the maximum information principle. *Educational and Psychological Measurement*, **57**, 2, 210-221.

Reckase M.D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Statistics*, 4, 207-230.

Renom J. (1993). *Tests adaptativos computerizados. Fundamentos y aplicaciones*. Barcelona: PPU.

EJERCICIOS

1. Dos ítems han sido calibrados con el modelo logístico de dos parámetros. Sus parámetros son a_1=1 y b_1= 0.5, y a_2= 0.5 y b_2= 2.5.

a) ¿Qué item de los dos es más fácil?
b) ¿Qué item es más discriminativo?
c) Una persona cuya θ= 2 responde a los dos ítems (y se cumple el supuesto de independencia local) ¿Cual es la probabilidad de que falle los dos? ¿Cual la de que acierte los dos? ¿Cual la de que acierte uno y falle el otro?

2. Aplicamos tres ítems a 5 personas y sus respuestas han sido las siguientes (1, acierto; 0, error):

Persona	Ítem 1	Ítem 2	Ítem 3
1	1	0	1
2	1	1	0
3	1	1	0
4	1	0	0
5	0	1	1

Procedemos a la estimación conjunta, mediante el modelo L3P, de los parámetros de los ítems y de los niveles de rasgo de las cinco personas.

a) Todas las personas menos la número cuatro tienen dos aciertos, por lo tanto todas ellas menos la cuatro deberán obtener el mismo nivel de rasgo. V( ) F( )
b) La dificultad del ítem 1, b_1, deberá de ser menor que la del ítem 2, b_2. V( ) F( ).
c) El parámetro "c" deberá ser 1/3, pues sólo hay tres ítems. V( ) F( ).

3. Obtenga cuanto vale la probabilidad de acertar un ítem en el modelo logístico de tres parámetros cuando el nivel del habilidad de la persona coincide con la dificultad del item.

4. Los tres parámetros de un ítem son a= 2, b= 1 y c= 0.2.

a) ¿Qué probabilidad de acertar el ítem tiene una persona con nivel de habilidad θ = 0?
b) ¿Puede corresponder a una persona una probabilidad de acertar de 0.10 en ese ítem?

5. Elegimos cien personas que tienen exactamente el mismo nivel de rasgo. Se les pasa un ítem fácil (b= -1) y lo aciertan 80 de los cien. Se les pasa a continuación un ítem más difícil (b= 0.5) y lo aciertan 40 de los cien. Supongamos que se cumplen los supuestos de la TRI ¿cabe esperar que los 40 que han resuelto el segundo ítem, el más difícil, hayan también resuelto el ítem más fácil?

6. La θ estimada de Andrés es 1.2 y su error típico de estimación 0.15.

a) Obtenga el intervalo en el que se encuentra la puntuación θ de Andrés con probabilidad 0.99.

b) Obtenga la θ estimada y el error típico de estimación de Antonio, sabiendo que su θ está entre 1.12 y 2.10, con probabilidad 0.95.

7. Pasados varios ítems, un TAI estima a una persona una θ de 0.5. Los parámetros de dos ítems que todavía no han sido aplicados son los siguientes: a_1= 1, b_1= 0.5, a_2= 2 y b_2= 1. Si el TAI ha de suministrar un ítem de estos dos ¿Cual suministraría? (La derivada P' correspondiente al modelo logístico de dos parámetros es P'= DaPQ).

8. El nivel de rasgo de una persona es 1 (es decir, θ= 1). Aplicando el modelo logístico de dos parámetros obtenemos las CCIs de dos items. La probabilidad de acierto en el item 1 (a_1= 1 y b_1= 0.5) coincide con su probabilidad de acierto en el item 2 (a_2= 2, b_2 no conocido). Obtenga cuanto vale b_2.

9. Tenemos tres ítems. La probabilidad de acierto, en cada uno de ellos, que corresponde a cada uno de los siguientes valores de θ se ofrece a continuación:

	P(θ)						
	-3	-2	-1	0	1	2	3
Ítem 1	0.1	0.15	0.2	0.35	0.5	0.65	0.8
Ítem 2	0.0	0.0	0.0	0.10	0.5	0.90	1.0
Ítem 3	0.0	0.10	0.5	0.90	1.0	1.0	1.0

a) Dibuje las tres CCIs.
b) Compare la dificultad y poder discriminativo de los ítems 1 y 2. ¿Cual es más difícil? ¿Cual es más discriminativo?
c) Compare la dificultad y poder discriminativo de los ítems 2 y 3. ¿Cual es más difícil? ¿Cual es más discriminativo?

10. Un test consta de sólo dos ítems (a_1=1, b_1=0, y a_2=2, b_2=-1).

a) Obtenga la función de información del test para los valores de θ= -3, -2, -1, 0, 1, 2 y 3. (Se recuerda que en el modelo logístico de dos parámetros P'= DaPQ).
b) ¿Para qué valor de θ (de los expuestos anteriormente) el test proporciona la máxima información?
c) Obtenga el error típico de estimación con el que el test estimaría la θ de una persona cuya θ real fuese -1.

SOLUCIONES

1. a) El ítem 1, pues $b_1 < b_2$.
b) El ítem 1, pues $a_1 > a_2$.
c) 0.367 (dos aciertos), 0.043 (dos fallos) y 0.590 (un acierto y un fallo).

2. F, V, F.

3. (1+c)/2.

4. a) P(0) = 0.226.
b) Es imposible, pues $P(\theta) \geq c = 0.20$.

5. No. La independencia local supone que la probabilidad de acertar ambos items será (80/100)(40/100) = 0.32.

6. a) límite inferior: 1.2-(2.56)(0.15) = 0.816.
límite superior: 1.2+(2.56)(0.15) = 1.584.
b) zeta estimada = (1.12+2.10)/2 = 1.61.
error típico = (2.10-1.61)/1.96 = 0.25.

7. En el primer ítem, P=0.5, Q=0.5, P'=0.425 e I(0.5)=0.72.
En el segundo, P=0.15, Q=0.85, P'=0.446 e I(0.5)=1.47.
El TAI aplicaría el segundo ítem, a pesar de que la dificultad del primero coincide con la θ estimada.

8. Ha de cumplirse que $a_1/a_2 = (\theta-b_2)/(\theta-b_1)$.
De donde, b_2= 0.75.

9. b) Misma dificultad ($b_1=b_2=1$). Más discriminativo, el dos; pues su CCI tiene más pendiente.
c) Más fácil el ítem tres (b_3= -1) y misma discriminación.

10. a) En ítem 1, I(-3)= 0.017, I(-2)= 0.091, I(-1)=0.376, I(0)= 0.72, I(1)= 0.376, I(2)=0.091 e I(3)=0.017.
En ítem 2, I(-3)= 0.014, I(-2)= 0.364, I(-1)=2.89, I(0)= 0.364, I(1)= 0.014, I(2)=0.000 e I(3)=0.000.
En el test, I(-3)= 0.031, I(-2)= 0.455, I(-1)=3.266, I(0)= 1.084, I(1)= 0.390, I(2)=0.091 e I(3)=0.017.
b) De los niveles de rasgo considerados, el que se estimaría con mayor precisión es θ = -1. Es decir, el test resulta máximamente informativo para θ= -1.
c) $S^2e = 1/3.266 = 0.306$. Se = 0.55.