0,65 €

TEM 4
Psicometría
José Antonio Pérez Gil

TEMA 4

Evaluación del instrumento de medida: Análisis de ítems.

4.1. Introducción.

4.2. Análisis de la calidad de los ítems.
- 4.2.1. Información descriptiva: validez de los ítems.
- 4.2.2. Información estadística: índices estadísticos de los ítems.
  - 4.2.2.1. Índice de Dificultad
  - 4.2.2.2. Índice de Discriminación
  - 4.2.2.3. Índice de Validez
- 4.2.3. Funcionamiento Diferencial de los Ítems (FDI).

4.1.- Introducción.

En la construcción de un test el objetivo general es llegar a un test de mínima longitud que produzca puntuaciones con el mayor grado posible de fiabilidad y validez. Esto se lleva a cabo partiendo de un número considerable de ítems (2, 3 o, incluso, 4 veces más ítems que el test resultante) que se someten, en primer lugar, al juicio de expertos, según hemos apuntado en el capítulo anterior. Una vez depurados estos ítems y reducidos a un menor número de ellos se pasan a una o varias muestras de sujetos en lo que se suele llamar estudio piloto, para realizar con los datos así obtenidos lo que se suele llamar análisis de ítems. Este proceso consiste en estimar las propiedades estadísticas de los ítems que más van a influir en la fiabilidad y validez del test resultante. Pues cuanto mejores sean los ítems mejor será el test que resulte de la unión de ellos.

La fiabilidad y la validez son las dos propiedades fundamentales de un buen instrumento de medida; son propiedades exigibles a un buen test y, por lo tanto, criterios a tener en cuenta para depurar los ítems.

En este capítulo vamos a describir los procedimientos y técnicas para contrastar tanto la calidad de los ítems, como la fiabilidad y validez del instrumento construido. Veamos estos aspectos.

4.2. Análisis de la calidad de los ítems.

Si hemos seguido los pasos recomendados en los apartados anteriores de este capítulo, posiblemente vayamos por buen camino en la construcción del instrumento de medida. Pero, a veces el autor del test tiene su juicio viciado por la cercanía e implicación que tiene con la prueba. Llega el momento de contrastar por procedimientos distintos y lo más objetivos posible la bondad del ítem, que propiciará de forma inexcusable la consecuente bondad de la prueba en su conjunto.

No basta con evaluar la calidad global de la prueba mediante indicadores de su fiabilidad y validez. Es menester también proceder al análisis de los elementos que componen dicha prueba, esto es, al examen de la calidad individual de cada uno de esos elementos. En particular, el análisis de ítems supone la obtención de información descriptiva y estadística.

La **información descriptiva** hace e referencia a la calidad técnica de los elementos y en su obtención suele ser muy frecuente el recurso a un grupo de jueces o expertos. Habitualmente, cada juez emite un juicio en relación a la relevancia, adecuación, claridad, etc. en la redacción del enunciado del ítem y de sus opciones de respuesta, si las hay; se concede una valoración positiva a un ítem cuando al menos un elevado porcentaje de jueces lo valoran de ese modo.

La **información estadística** supone la estimación de una serie de parámetros para cada ítem, el análisis de los distractores o alternativas incorrectas de respuestas -en los ítems de

elección múltiple- y el examen del posible funcionamiento diferencial de los ítems en grupos de interés.

Dentro de los **parámetros de los ítems** se suele estimar la dificultad, discriminación, pseudoadivinación, homogeneidad, información, fiabilidad y validez, si bien según sea el marco teórico que se adopte teoría clásica de los tests, teoría de respuesta al ítem o medición referida al criterio- algunos de ellos serán definidos e interpretados/valorados de distinto modo y se estimará un determinado número de estos parámetros, no todos. La mayor parte de los textos de psicometría presentan algún capítulo dedicado a estas cuestiones (véase, por ejemplo, Crocker y Algina, 1986; Martínez Arias, 1995; Muñiz, 1992 o Santisteban, 1990).

No puede decirse otro tanto en relación al tema del **análisis de los distractores**, que ha recibido un tratamiento considerablemente más breve en la literatura, a pesar de que el distractor constituye una parte -y una parte importante- del elemento. Del mismo modo que se analizan los ítems y, a continuación, se procede a su revisión, sustitución o supresión, hay que detectar qué distractores funcionan en la dirección esperada y cuáles no, y proceder subsiguientemente a su revisión, sustitución o supresión. Haladyna (1994) propone tres vías para recoger información acerca del funcionamiento de los distractores:

1. La tabla de frecuencias (Levine y Drasgow, 1982; Walner, 1989).
2. La curva característica del distractor (Thissen, Steinberg y Fitzpatrick, 1989).
3. Índices estadísticos (correlación distractor-puntuación total en el test, media en el test de los sujetos que eligen un distractor, χ^2).

En resumen, los principales criterios de bondad de una prueba y subsecuentemente de un ítem son dos: validez y fiabilidad.

Para contrastar la calidad técnica de los ítems Osterlind, (1989), recomienda dos procedimientos fundamentales en los que él denomina el *juicio de expertos*: El juicio de expertos, como procedimiento, permite obtener validez de contenido para cada uno de los ítems del instrumento y por tanto, del propio instrumento de medida globalmente.

En resumen, el examen de la calidad individual de cada uno de los elementos, ítems o indicadores que componen un instrumento de medida implica la obtención de información que denominaremos descriptiva y estadística.

- Información descriptiva: hace e referencia a la calidad técnica de sus elementos y en general supone la obtención de evidencias de validez de los mismos, es decir, obtención de indicios sobre su relevancia, adecuación, claridad, etc.
- Información estadística: supone la estimación de una serie de parámetros para cada ítem, el análisis de los distractores o alternativas incorrectas de respuestas -en los ítems de elección múltiple- y el examen del posible funcionamiento diferencial de los ítems en grupos de interés.

AFRIL

4.2.1. Información descriptiva: validez de los ítems.

El objetivo de este apartado es establecer juicios razonables sobre el grado en que las evidencias aportadas permiten emprender acciones basadas en el modelo de medida desarrollado, en este caso, a través de los ítems o indicadores (Messick, 1989). No cabe plantear si el sistema de ítems o indicadores es válido en sí mismo independientemente del objetivo de la investigación o del contexto, sino que su validez está en relación con los objetivos a medir.

No hay un coeficiente único que reporte datos sobre la validez de los ítems o sistema de indicadores, sino que los datos para establecer juicios sobre su validez vienen dados por:

- Los resultados de combinar distintas técnicas

- Por evidencias empíricas basadas en experiencias anteriores.

- La utilidad que muestren para el modelo de medida propuesto

De acuerdo con ello, es necesario dotar a los ítems o indicadores de distintas evidencias de validez, en concreto:

1. Validez de contenido: Garantizar que los indicadores seleccionados constituyan una muestra representativa de todos los posibles indicadores.
2. Validez de constructo: Garantizar la existencia del constructo que conforma el conjunto de indicadores que pretenden medirlo y por ende dota de sentido a las puntuaciones que se obtienen con los indicadores.
3. Validez de criterio: Conjunto de evidencias que permiten demostrar que las puntuaciones del sistema de indicadores están relacionadas con un criterio externo de interés (Suen, 1990).

En general, los distintos tipos de validez, no son entidades distintas, sino que son el resultado de distintas aproximaciones al concepto de validez de los indicadores o ítems en nuestro caso, con criterios de utilidad para nuestros propósitos, de correspondencia con el modelo teórico referente y de manera que los ítems o el sistema de indicadores sea coherente con investigaciones previas. Todo ello, nos permitirá obtener un Modelo de medida a partir del cual programar acciones en el campo aplicado, en nuestro caso, la psicología.

Las evidencias de Validez de contenido nos permitirán garantizar que los indicadores seleccionados constituyan una muestra representativa de todos los posibles indicadores.

El *juicio de expertos* para contrastar la validez de los ítems consiste en preguntar a personas expertas en el dominio que miden los ítems, sobre su grado de adecuación a un criterio determinado y previamente establecido en los pasos anteriormente reseñados de construcción de una prueba. Para esta validación se siguen dos procedimientos (Osterlind, 1989):

AFRIL

A) Juicios del grado de adecuación entre cada ítem y el subdominio u objetivo que pretende medir el ítem. Ver la tabla 4.1. como ejemplo de hoja de respuesta para los jueces.

Nombre_____

Instrucciones: Primero lea el objetivo número 1. Después, lea el primer ítem en el cuadernillo del test. Considere cuidadosamente el grado de congruencia de este ítem con el dominio de habilidad. Evalúe la congruencia según las siguientes categorías:

A = Alto grado de congruencia
M = Grado medio de congruencia
B = Bajo grado de congruencia

Si Vd. Quiere hacer algún comentario sobre la congruencia de ese ítem regístrelo en el lugar reservado para ello. Después de que haya terminado con el primer ítem, proceda con el segundo ítem, e igualmente con todos los ítems siguientes, evaluándolos de igual forma.

Objetivo Nº 1
Descripción del contenido sustantivo del objetivo n° 1.

	Evaluación	**Comentario**
Ítem Nº 1		
Ítem Nº 3		
Ítem Nº 7		
Ítem Nº 9		
Ítem Nº 10		

Objetivo Nº 2
Descripción del contenido sustantivo del objetivo n° 2.

	Evaluación	**Comentario**
Ítem Nº 2		
Ítem Nº 4		
Ítem Nº 5		
Ítem Nº 6		
Ítem Nº 8		
Ítem Nº 11		

Tabla 4.1. Grado de congruencia ítem-objetivo (adaptado de Osterlind, 1989, p.267).

Si no nos basta con la simple cuantificación del número de jueces que están de acuerdo, podemos recurrir al índice de congruencia entre el ítem **"i"** y el objetivo **"k"** (Rovinelli y Hambleton, 1977; Hambleton, 1980):

$$I_{ik} = \frac{(N-1)\sum_{j=1}^{N} X_{ijk} + N\sum_{j=1}^{n} X_{ijk} - \sum_{j=1}^{n} X_{ijk}}{2(N-1)n} \qquad (4.1)$$

ATRIL

Por ejemplo (Osterlind, op. cit. p. 270), en un test con 36 ítems y 5 objetivos específicos se evalúa la congruencia entre el ítem 1 con el objetivo 2. Nueve expertos han emitido un juicio: uno opina que no hay unión entre el ítem y el objetivo, otro que la unión es débil y siete que la unión es fuerte. Así, la suma de los nueve juicios es -1+0+(1*7) = 6. aplicando el ejemplo a la ecuación 4.1.

$$I_{21} = \frac{(5-1)6 + 5.6 - 6}{2(5-1)9} = 0.67 \qquad (4.2)$$

B) *Emparejamiento de ítems y objetivo/s o dominio/s*, a juicio de los jueces.- En este procedimiento se oculta a los jueces el emparejamiento objetivo-ítem para comprobar si, dicho emparejamiento, es tan evidente que una persona experta lo pueda descubrir. Ver tabla 4.2. como ejemplo del protocolo de respuesta. En este caso se puede utilizar la proporción de aciertos como índice de congruencia.

Marcas de emparejamiento entre ítems y objetivos

Juez: _____

Fecha: / /

ÁREA de contenido: _____

Primero lea detenidamente toda la lista de especificaciones de dominio y de ítems del test. Su marca es para señalar qué ítem del test es y cuál no es una medida de una de las especificaciones de dominio. Esto es, si Ud. Cree reconocer dentro de un banco de ítems la representación en un ítem del test de la medida de una especificación de dominio, puede emparejarlo con una linea al objetivo correspondiente. Esto es, después de cada objetivo, escribir el/los ítems del test que crea miden dicho objetivo. En el caso de aquellos ítems que Vd. crea que no miden ninguna de las especificaciones de dominio disponibles, marcar sus números en el espacio disponible al final.

Objetivo	Emparejamiento de ítems del test.
1	
2	
3	
4	

Ítems no emparejados:

_____

From R.K. Hambleton "Validating the test scores" (p.225) in R.A. Berk (ed), *A Guide to Criterion-Referenced Test Construction*, 1984, Baltimore: The Johns Hopkins University Press

Tabla 4.2. Hoja resumen de las señales de emparejamiento de los jueces entre objetivos e ítems (adaptada de Osterlind, 1989, p. 268).

AFRL

Las evidencias de Validez de constructo nos permitirán garantizar la existencia del constructo que conforma el conjunto de indicadores que pretenden medirlo y por ende dota de sentido a las puntuaciones que se obtienen con los indicadores.

Para contrastar la calidad técnica de los ítems, en cuanto a su validez de constructo suele recurrirse al estudios del modelo de medida desde la técnica del Análisis Factorial (Exploratorio y Confirmatorio).

En general, el Análisis Factorial (AF) como modelo de medida asume que los factores son causas efectivas de los indicadores, que son vistos como efectos de los mismos. Los indicadores o ítems son efectos (manifestaciones) de las variables latentes o conceptos psicológicos no observables.

Las evidencias de Validez de criterio se refieren a un conjunto de evidencias que permiten demostrar que las puntuaciones del sistema de indicadores están relacionadas con un criterio externo de interés (Suen, 1990).

Para contrastar la calidad técnica de los ítems desde la perspectiva de validez de criterio, se suelen utilizar procedimientos que implican evaluar el poder predictivo del instrumento globalmente. En este sentido nos remitimos al concepto de validez del instrumento que analizaremos en apartados subsiguientes.

4.2.2. Información estadística: índices estadísticos de los ítems.

Dentro del marco de la TCT, los parámetros de los ítems que se suelen estimar son el índice de dificultad, discriminación y validez, si bien según sea el marco teórico que se adopte TCT, TRI o TG se valoran otros índices (homogeneidad, información, pseudoadivinación...). Algunos de ellos serán definidos e interpretados/valorados de distinto modo según el marco téorico.

La mayor parte de los textos de psicometría presentan algún capítulo dedicado a estas cuestiones (véase, por ejemplo, Crocker y Algina, 1986; Martínez Arias, 1995; Muñiz, 1992/98 o Santisteban, 1990).

Como ya hemos apuntado, en la construcción de un test el objetivo general es llegar a un test de mínima longitud que produzca puntuaciones con el mayor grado posible de fiabilidad y validez. Esto se realiza, según hemos apuntado en el apartado anterior, sometiéndolos, en primer lugar, a juicio de expertos. Una vez depurados estos ítems y reducidos a un menor número de ellos se pasan a una o varias muestras de sujetos en lo que se suele llamar estudio piloto, para realizar con los datos así obtenidos lo que se suele llamar análisis de ítems. Este proceso consiste en estimar las propiedades estadísticas de los ítems que más van a influir en la fiabilidad y

validez del test resultante. Pues cuanto mejores sean los ítems mejor será el test que resulte de la unión de ellos.

De acuerdo con lo dicho en el párrafo anterior los parámetros que se suelen analizar en los ítems de cara a su consiguiente depuración se pueden resumir en tres grupos (Gulliksen, 1950; Croker y Algina, 1986; Osterlind, 1989; Martínez-Arias, 1995; Muñiz, 1994): Índices de *dificultad*, *discriminación* y *validez*. (veamos a continuación cada uno de ellos).

4..2.2.1. Índice de Dificultad

Como entrada cabe afirmar que, en igualdad para el resto de condiciones, un ítem es mejor cuanto mayor sea su varianza.

De lo dicho en el párrafo anterior se desprende que un ítem que se evalúa de forma dicotómica –correcto/incorrecto- tiene máxima varianza cuando p=q=0.5. Ya que es entonces cuando su varianza (que es igual a p*q) es máxima. Por lo tanto este es un índice importante y a tener en cuanta a la hora de seleccionar los ítems. Nos estamos refiriendo a lo que se suele llamar **índice de dificultad**, que no es otra cosa que la proporción de sujetos que responden correctamente al ítem. Éste es un índice descriptivo de la distribución de respuestas y/o puntuaciones del ítem:

$$ID = \frac{A}{N} \qquad (4.3)$$

donde:

A: Número de sujetos que aciertan el ítem.

N: Número de sujetos que han intentado resolver el ítem.

En los ítems de respuesta limitada a un número delimitado de alternativas, conviene corregir el número de sujetos que han acertado el ítem (*A*), restándole el número de ellos que lo han acertado por azar (*E/(K-1)*, para así obtener los que en realidad saben el ítem. El proceso de razonamiento es análogo al que se sigue para demostrar la fórmula 6.10.

$$ID = \frac{A - \frac{E}{(K-1)}}{N} \qquad (4.4)$$

Donde:

E: Numero de sujetos que han fallado el ítem
K: Número de alternativas de respuesta del ítem

Como ya se ha apuntado anteriormente, los mejores ítems son los que tienen un índice de dificultad de 0.5, porque optimizan la variabilidad del ítem[1]. Por lo que respecta al test en su

[1] En los tests de personalidad o actitudes no cabe hablar de índice de dificultad, pero todo lo que hemos dicho del ID. cabe afirmarlo respecto de la proporción de sujetos que saturan con el atributo, esto es, que responden en el sentido de manifestar el constructo que estamos midiendo.

conjunto, para que discrimine no sólo entre los de aptitudes intermedias (como ocurre si todos los ítems son de dificultad intermedia), sino también entre los de niveles de habilidad más extremos. Yela (1980) recomienda la siguiente distribución de ítems en el test (tabla 4.3):

Categorías	Centil en el constructo	% Aproximado	Índices de Dificultad.
Muy fáciles	*0 - 10*	*10%*	*0.75 a 0.95*
Fáciles	*10 – 30*	*20%*	*0.55 a 0.74*
Normales	*30 - 70*	*40%*	*0.45 a 0.54*
Difíciles	*70 - 90*	*20%*	*0.25 a 0.44*
Muy fáciles	*90 – 100*	*10%*	*0.05 a 0.24*

Tabla 4.3. Distribución de ítems por ID en un test general. (Yela, 1980).

En este mismo sentido, Garret (1968) propone (tabla 4.4.):

Categorías	Centil en el constructo	% Aproximado	Índices de Dificultad.
Fáciles	*0 - 25*	*25%*	*0.75 a 0.95*
Medios	*25 - 75*	*50%*	*0.26 a 0.74*
Difíciles	*75 – 100*	*25%*	*0.05 a 0.25*

Tabla 4.4. Distribución de ítems por ID en un test general Garret (1968) .

Osterlind recomienda, en tests de rendimiento, índices de dificultad que oscilen entre 0.4 y 0.8. También considera que puede ser útil analizar la evolución del índice de dificultad de un ítem para subpoblaciones con distintos niveles en el constructo (ver figura 4.1.).

Si, al contrario de lo que se observa en el gráfico representado en fig. 4.1., no se apreciara evolución o cambio en los ID de un grupo a otro sería un síntoma de falta de fiabilidad y/o de validez del ítem. Este problema quedaría solventado con tests a la medida de la cantidad del constructo que posee cada sujeto. Dado que el índice de dificultad depende de la población en la que se calcula, con los tests a medida todos los ID. serían siempre de 0.5.

Como es lógico, solo cabe hablar de índice de dificultad en los tests de aptitud máxima. Sin embargo, para cualquier tipo de prueba psicológica en la que los ítems ofrezcan varias alternativas de respuesta, puede resultar útil analizar la distribución de frecuencias para cada una de las alternativas de respuesta. Siempre cabe afirmar que una alternativa que no la elija nadie o casi nadie es inútil. A este respecto, en tests de aptitud máxima, lo lógico es que:

- La alternativa correcta sea elegida mayor número de veces que las incorrectas.
- Los distractores o alternativas incorrectas sean elegidas todas ellas aproximadamente igual número de veces.

AFRIL

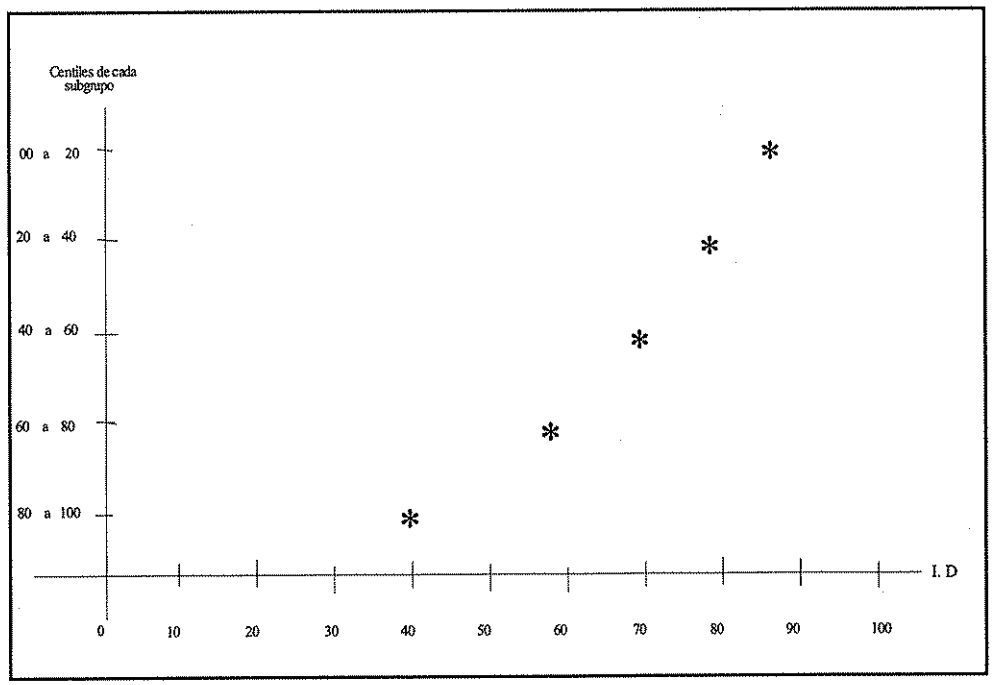

Figura 4.1. ID de un ítem por niveles de habilidad.

Con lo dicho sobre el ID, ampliado con otras lecturas, más el sentido común, podemos tomar las decisiones más acertadas en función de nuestras necesidades y objetivos de medida. En este sentido, si tenemos que seleccionar al 30% de aspirantes más hábiles, nos interesan ítems con ID=0.3. En tests con ítems homogéneos pueden interesarnos niveles de dificultad que cubran todo el rango (ver cuadros orientativos anteriores de Yela (1980) y Garret (1968)). En tests heterogéneos puede interesarnos que todos o la mayoría de los ítems tengan valores centrales (alrededor de 0.5) de dificultad. En tests referidos al criterio el ID será el adecuado al punto de corte.

Hay un problema asociado al ID como es el hecho de que, en los tests de varias alternativas de respuesta, los sujetos no responden con un grado de certeza total o nulo. Este tema ha sido tratado por distintos autores sin haberse encontrado una solución plenamente satisfactoria, aunque sí se han ofrecido diferentes alternativas:

- Juicios de seguridad por parte de los evaluados.
- Responder hasta acertar y al puntuar penalizar el número de errores.
- Ponderar las alternativas de los ítems, en función del juicio de expertos sobre el grado de corrección o incorrección de cada una de ellas.
- Tener en cuenta la tabla adjunta (ver tabla 4.5) a la hora de calcular el ID o la puntuación directa de un sujeto, siguiendo la fórmula $P_{óptimo} = P + (1-P)/k$

AFRL

K	Proporción que conoce la respuesta	Proporción que acierta por azar	Proporción total Po	Po de Lord[2]
4	0.50	0.50/4	0.62	0.74
3	0.50	0.50/3	0.67	0.77
2	0.50	0.50/2	0.75	0.85

Tabla 4.5. Valores óptimos de proporción (P_o) según el número de alternativas. (adaptado de Martínez-Arias, 1995).

Esta distribución de frecuencias de las distintas alternativas de respuesta la podemos hacer también dividiendo la muestra en dos grupos extremos por lo que respecta a su rendimiento en el test completo o en cualquier otro criterio externo. A partir de la tabla resultante (ver tabla 4.6) se obtiene información de gran interés por lo que respecta al llamado análisis de alternativa incorrectas o, en general, a la relación entre alternativa de respuesta elegida y pertenencia al grupo superior o inferior en el test o en un criterio (ejemplo tomado de Muñiz, 1998; p.230). En esta ocasión la alternativa E es inútil porque no la elige nadie y la D parece ser que ofrece información que confunde precisamente a los del grupo superior. También habría que revisar la alternativa B, por ser especialmente atractiva.

		Alternativas					
		A	*B*	*C**	*D*	*E*	
Ítem	*Superior*	*5*	*15*	*70*	*10*	*0*	*100*
	Inferior	*15*	*65*	*20*	*0*	*0*	*100*
		20	*80*	*90*	*10*	*0*	*200*

Tabla 4.6. Análisis de las alternativas incorrectas (Muñiz, 1998).

La información obtenida con este cuadro será retomada en el apartado siguiente, al hablar del índice de discriminación. En un apartado posterior sobre aplicaciones informáticas trataremos también, a través del programa MicroCAT, este aspecto del estudio de las distintas alternativas de respuesta y su relación con la puntuación total del test.

4.2.2.2. Índice de Discriminación.

El objetivo de la mayoría de los tests es ofrecer información acerca de las diferencias individuales en el constructo o en cualquier criterio externo cuya predicción, selección u optimización se pretende conseguir a través de las puntuaciones obtenidas en el test. Para la mejor consecución de ambos objetivos, un parámetro de gran interés en la selección de los ítems es aquel que ponga de manifiesto su poder para discriminar, esto es, diferenciar a los sujetos con puntuaciones altas en el test de aquellos que las tienen bajas. En definitiva, el índice de discriminación de un ítem es un índice que relaciona el ítem con el test y se define como *"la correlación entre las puntuaciones de los sujetos en el ítem y sus puntuaciones en el test"* (Muñiz, 1998, p.219). En cualquier caso, dado que el puntaje en el test suele ser la suma de la

[2] Lord (1952) afirma que la probabilidad de acertar al azar es mayor que 1/k debido al hecho del posible conocimiento parcial de la respuesta.

puntuación en los ítems, hay que evitar la influencia en el índice obtenido de la participación del ítem en la puntuación total del test (para mayor información se puede consultar Muñiz, 1998 y Martínez-Arias, 1995).

- Índice de discriminación clásico (***D***) propiamente dicho (Croker y Algina, 1986):

$$D = Ps - Pi \tag{4.5}$$

donde:

Ps es la proporción de sujetos del grupo superior[3] en el criterio que responden correctamente al ítem. y

Pi es la proporción de sujetos del grupo inferior que responden correctamente al ítem.

Distintos tipos de correlaciones entre el ítem y el test para calcular el índice de discriminación:

- Coeficiente de Correlación producto-momento de Pearson ($\mathbf{r_{xy}}$)

$$r_{xy} = \frac{\frac{\sum xy}{N} - \overline{X} * \overline{Y}}{S_x * S_y} \tag{4.6}$$

Aplicaciones o estimaciones del Coeficiente de Correlación de Pearson (casos particulares):

- Correlación Biserial Puntual (ρ_{bp})

$$\rho_{bp} = \frac{\mu_p - \mu_x}{\sigma_x}\sqrt{\frac{p}{q}} \tag{4.7}$$

donde:

μ_p: Media en el test de los sujetos que han acertado el ítem

μ_x: Media del test

σ_x: Desviación típica del test

p: Proporción de sujetos que aciertan el ítem

q: 1-p

- Correlación Biserial (ρ_b)

$$\rho_b = \frac{\mu_p - \mu_x}{\sigma_x} * \frac{p}{y} \tag{4.8}$$

donde:

y: Ordenada correspondiente al valor de la puntuación típica en la curva normal que deja por debajo un área igual a "p" (los valores se pueden encontrar en la tabla estadística

[3] Recordemos la formación de un grupo superior y otro inferior en tabla 7.6.

correspondiente). Los demás elementos tienen el mismo significado que en la expresión 4.7.

- Correlación Phi (φ)

$$\Phi = \frac{bc - ad}{\sqrt{(a+b)(b+c)(c+d)(b+d}} \qquad (4.9)$$

donde:

a,b,c, y ***d*** son las frecuencias de cada una de las cuatro casillas formadas por el cruce de un ítem dicotómico y un tests dicotómico.

- Correlación Tetracórica (ρ_t)

$$\rho_t = \cos\frac{180}{1+\sqrt{\frac{bc}{ad}}} \qquad (4.10)$$

donde:

a,b,c, y ***d*** son las frecuencias de cada una de las cuatro casillas formadas por el cruce de un ítem y un test ambos dicotomizados. asumiendo distribuciones normales

Hasta aquí, en este apartado del índice de discriminación, nos hemos limitado a presentar distintos procedimientos estadísticos para la estimación de dicho parámetro. Nuestro interés irá dirigido al adecuado uso e interpretación de estos índices, más que a su cálculo, que es un problema ya resuelto por la informática. Generalizando, podemos decir que, en todos los casos, cuanto mayor sea el índice de discriminación o de correlación del ítem con el criterio mejor es el ítem.

Resumiendo, vamos a dar unas pautas generales de uso e interpretación de los distintos estimadores de la discriminación de un ítem:

- Con ítems de dificultad intermedia el Índice de Discriminación **D** es fácil de calcular y de interpretar (oscila entre -1 y +1, y lo normal es que sea positivo). Tiene el inconveniente de que se ve muy afectado en el caso de los ítems con índices de dificultad extremos. Así, p.e., para un ítem con **ID**=0.9 el valor máximo de **D**=0.2 . Otro inconveniente de este índice es que no tiene test de significación estadística.
- Si los ítems son de dificultad extrema es adecuado utilizar ρ_b por estar poco afectado tanto por ID como por la fluctuación muestral. Tiene la exigencia de que ambas distribuciones (las de test e ítem) sean normales (la del ítem dicotomizada).
- Si se sospecha que distintas muestras de la población serán similares en habilidad, y si el objetivo es seleccionar ítems con alta consistencia interna; lo adecuado es

AFRIL

utilizar ρ_{bp} (Lord y Novick, 1968). Tiene el inconveniente de que los ítems han de ser dicotómicos.

- Cuando ítem y test son dicotomizados y con distribuciones normales en origen, ϕ y ρ_t son los más utilizados. ϕ es fácil de calcular pero está restringido cuando las proporciones de dicotomizaciones no son iguales. En este caso, sólo disponemos como última instancia de ρ_t , que además es el coeficiente de correlación más utilizado para someter posteriormente la matriz de correlaciones a análisis factorial.

Hasta ahora hemos hablado de índices descriptivos y de índices correlacionales. Como ambos tipos de índices es pertinente que sean tenidos en cuenta, presentamos a continuación el índice de validez.

4.2.2.3. Índice de Validez

El índice de validez de un ítems está referido a su correlación con un criterio externo al propio test. Como ocurre con el índice de discriminación su calculo dependerá de la naturaleza de las variables con las que se correlacione el ítem, es decir, las distintas fórmulas de correlación reseñadas en el índice de discriminación pueden ser utilizadas para el cálculo del índice de validez. Su cálculo es idéntico, si bien ahora no existe el problema adicional de que el ítem esté incluido en el criterio como podía ocurrir en el caso del índice de discriminación.

A modo de conclusión o resumen.

Después de tantos índices: de dificultad, correlacionales y validez, ¿qué cabe decir de todos ellos a modo de resumen o conclusión? Pues caben algunas reflexiones, pero la principal, y por encima de todas ellas, es: tener siempre presente el objetivo final para el que se quiere realizar el escalamiento, y subordinar a su consecución la información que nos proporcionen todos estos índices tanto de los ítems como del test.

Se podría decir que la consigna a la hora de interpretar estos índices, algunos de ellos contrarios o enfrentados, es aplicar el sentido común, sabiendo lo que informa cada índice y sabiendo lo que se quiere conseguir como objetivo de escalamiento. Pero, dado que el sentido común no siempre es igual para todos, vamos a hacer algunas recomendaciones, sin intentar ser exhaustivos ni excesivamente tajantes.

La virtud principal que ha de cumplir cualquier prueba es la validez. Si una prueba se demuestra que es altamente válida, consiguientemente y como condición necesaria, ha de ser fiable. Porque de no ser así, la fluctuación o azar de las distintas medidas, podría afectar a la relación o explicación mutua entre los constructos y evitaría que se pusiera de manifiesto en los distintos estudios empíricos de validez como veremos más adelante. Por lo tanto, si se consigue validez nos podremos quedar satisfechos. Pero como la validez es tan difícil de contrastar en algunos casos, se suele aconsejar optimizar la fiabilidad para así disponer, al menos, de la condición necesaria para, en la medida de lo posible, conseguir la validez.

AFRIL

Hay casos en los que lo medido es un único factor o dimensión psicológica. En tales circunstancias el depurar los ítems en aras de su fiabilidad no perjudica la validez, ya que se puede llegar al extremo de que, al coincidir el constructo que se mide con el test y el que se pretende medir, fiabilidad y validez de constructo son una misma cosa. Así pues, en tales casos es importante asegurar la validez de contenido previamente, especialmente en los tests de rendimiento que tienen referentes operativos (recordar figura 3.1 y tabla 3.2. del tema 3). De cualquier forma, sea cual sea el tipo de test, su validez estará más garantizada si la relación del constructo que medimos con otros atributos psicológicos es coincidente con teorías psicológicas consistentes y previamente hipotetizadas. Todos estos temas serán tratados con mayor extensión en el tema específico de validez.

Seguimos con algunas reflexiones generales sobre los distintos estadísticos utilizados para depurar los ítems. En pruebas de rendimiento o de aptitud práctica, depurar excesivamente los ítems para mejorar su fiabilidad puede perjudicar la mayor validez de una prueba más heterogénea y completa (con mayor validez de contenido).

De cualquier forma, la depuración de los ítems es cíclica y, aunque en algún caso se pueda admitir que se ha conseguido un buen nivel, con el tiempo puede aconsejarse su revisión.

En definitiva, el proceso a seguir, normalmente, es hacer el análisis de ítems en distintas muestras. Una vez depurados los ítems, se pasan a una muestra distinta y se vuelven a analizar los resultados obtenidos. Cuando ya se suele admitir que están depurados se contrasta su validez, y, si no es aceptable, se vuelve a empezar.

Una vez descrito los distintos índices que solemos utilizar para valorar las virtudes estadísticas de los ítems, vamos a destacar brevemente la importancia del sesgo que pueden presentar los ítems, es decir, el problema que presentan cuando los ítem se comportan de modo diferente en sujetos o grupos de sujetos que supuestamente tienen el mismo valor en la variable medida.

7.2.3. Funcionamiento Diferencial de los Ítems (FDI).

El *Funcionamiento Diferencial del ítem* (FDI) es un término recientemente acuñado por Holland y Thayer (1988) y, de un modo muy general, se puede caracterizar como un indicador de la existencia en el proceso de medición de error sistemático respecto a ciertos grupos de la población que está siendo evaluada (Baron, 1988). Lo reciente del término no debe de inducir a error: si bien éste es un tema tratado tardíamente en la literatura psicométrica (Muñiz, 1990), es a finales de los años sesenta cuando empieza a recibir una atención considerable, aunque por aquel entonces la etiqueta utilizada no era FDI, sino sesgo. Como se verá, esta cuestión de etiquetado no es trivial.

Aunque pueden encontrarse antecedentes remotos en los trabajos de Binet, se suele citar como primer trabajo en la investigación del sesgo el de Ells, Davis, Havighurst, Herrick y Tyler (1951) que trataron de identificar el tipo de problemas o ítems en los que los niños de *status*

socioeconómico alto mostraban una actuación superior y el tipo de problemas que podían hacer bien niños de *status* socioeconómico bajo (sesgo cultural). La eclosión de trabajos en torno al tema a finales de la década de los sesenta y, sobre todo, a comienzos de la de los setenta se vincula a la aparición en los EE.UU. de un fuerte movimiento por los derechos civiles. Las denuncias sobre la parcialidad de los tests crecieron considerablemente al inicio de esa década; se alzaron voces airadas en defensa de los intereses de distintas minorías que acusaban a los tests de ser barreras para la igualdad social y de oportunidades económicas. De hecho, se fallaron sentencias judiciales importantes por discriminación en la selección de personal y en la admisión a instituciones educativas como la Hobson vs. Hansen (1967), Diana vs. State Board of Education (1970), Griggs vs. Duke Power Co. (1971), Larry P. vs. Riley (1979), además del pernicioso acuerdo conocido como la regla dorada que tanto eco y repercusión ha tenido en el campo de la evaluación. Como apunta Muñiz (1990), "*será la discusión social y psicológica teórica, ajena en gran parte al círculo psicométrico especializado, reactivada tras el polémico artículo de Jensen (1969), la que obligue a los psicómetras profesionales a generar esta metodología, para mostrar que sus tests no están sesgados, o sí.*" (pp. 121-122)

Por tanto, el gran interés social en el tratamiento igualitario de grupos o minorías étnicas y/o socio-políticas ha sido un factor determinante a la hora de estimular el estudio sobre el sesgo de los ítems y tests utilizados para la selección de personas o en la admisión, promoción y orientación educativas. El sesgo de los instrumentos de medida se ha revelado como algo más que una cuestión puramente técnica, convirtiéndose en una cuestión de debate público e incluso legal.

En estos momentos iniciales se considera el sesgo a nivel de test y no de ítem y la preocupación básica es dictaminar si los tests utilizados habitualmente en el ámbito educativo y empresarial -y construidos, normalmente, por la clase y raza económica y políticamente dominante- conducían a una discriminación en el acceso al mercado laboral y educativo de personas pertenecientes a clases y razas ni económica ni políticamente dominantes, esto es, si los tests estaban sesgados contra grupos minoritarios. La falacia igualitarista condujo a una definición equívoca del sesgo: "*si los tests hacen a los hombres, que son iguales, desiguales, los tests son injustos ya que la justicia está con la igualdad de oportunidades. Así, cualquier test en el que se encontrasen diferencias entre grupos étnicos, culturales o socioeconómicos, se consideraba injusto y sesgado.*" (Fidalgo, 1996, p.376).

Se propuso toda una batería de procedimientos de detección del sesgo en la selección o implementación del sesgo del test. Al mismo tiempo, los editores de tests estaban interesados en contar con procedimientos que permitieran identificar ítems problemáticos, para revisarlos y/o eliminarlos antes de publicar la versión final de la prueba (Scheuneman y Bleistein, 1989), comenzándose entonces a diseñar métodos para la detección del sesgo en el ítem (Berk, 1982; Rudner, Getson y Knight, 1980). Originariamente, se acuñó el término sesgo en el ítem para distinguirlo de los métodos de detección de sesgo en el test. Sin embargo, esta acepción resulta problemática, al tener distintas connotaciones para diferentes audiencias (Reynolds y Kamphaus,

AFIRL

1990). Así, desde el punto de vista legal, el término sesgo denota una práctica discriminatoria ilegal mientras que para el lego denota simplemente prejuicios; desde un punto de vista estadístico, el sesgo hace referencia a error sistemático en la estimación de un valor (como opuesto al error aleatorio) y, desde la perspectiva de la investigación psicométrica, este error sistemático ocurre en la determinación del valor escalar de un sujeto a partir de su puntuación en un test como resultado de su pertenencia a un determinado grupo (o de su valor o posición en alguna otra variable). Parece conveniente pues sustituir este término por otro semánticamente más neutral y que refleje de forma más precisa lo que se entiende técnicamente por sesgo (Fidalgo, 1996).

Aunque el acuerdo no es unánime, sí existe un consenso bastante generalizado entre los autores que investigan estas cuestiones en reservar el término funcionamiento diferencial del ítem para el análisis estadístico de la cuestión y el término sesgo para las inferencias sobre la naturaleza de las diferencias observadas (Scheuneman, 1982; Scheuneman y Bleistein, 1989).

En las líneas que sigue se tratará de proporcionar una definición -ahora sí- acotada del FDI, se presentarán rápidamente los métodos que se han utilizado en su detección y se comentarán brevemente las áreas más activas en la actualidad de la investigación sobre el tema, para finalizar con alguna consideración acerca de explicaciones plausibles del FDI, esto es, acerca del sesgo de los ítems. Para un tratamiento detallado de estas cuestiones, son de obligada referencia los trabajos de Berk (1982), Camilli y Shepard (1994), Cole y Moss (1989), Fidalgo (1996), Gómez e Hidalgo (1999), Holland y Wainer (1993), Jensen (1980), Millsap y Everson (1993), Osterlind (1983), Reynolds y Brown (1984), Rudner, Getson y Knight (1980) y Scheuneman y Bleistein (1989).

Se dice que un ítem presenta un funcionamiento diferencial cuando, para sujetos con idéntico nivel en la característica medida con el test, la probabilidad de un sujeto de responder correctamente al ítem depende del grupo al que éste pertenezca, no es la misma para sujetos de distintos grupos.

Cuando la probabilidad de responder correctamente, para sujetos con la misma habilidad estimada, es sistemáticamente mayor en un grupo que en otro a lo largo de todo el continuo de habilidad, se habla de FDI uniforme; en caso contrario, de FDI no uniforme.

Cuando la probabilidad de responder correctamente a un ítem es la misma para sujetos con idéntica habilidad estimada con independencia del grupo al que pertenezcan pero existen, asimismo, diferencias entre la actuación media en el ítem entre grupos distintos de sujetos se habla de impacto (Ackerman, 1992): el impacto revela una diferencia entre grupos en la actuación en un ítem causada por una diferencia real en la variable medida y no por fuentes sistemáticas de variación ajenas al constructo que mide el test.

Son numerosos los esfuerzos clasificatorios que recoge la literatura sobre los métodos de detección del FDI; véase, entre los más recientes, los de Fidalgo (1996), Martínez Arias (1995),

AFRIL

Millsap y Everson (1993) y Scheuneman y Bleistein (1989). La clasificación que aquí se propone se articula en tomo a cuatro grandes categorías (Navas, 1998):

i. métodos basados en la TCT: el método *delta* (Angoff, 1972, 1982; Angoff y Ford, 1973; Angoff y Sharon, 1974) y el método de discriminación (Green y Draper, 1972; Ozenne, van Gelder y Cohen, 1974).

Estos métodos han caído en total desuso al no ser métodos propiamente de detección de FDI sino simplemente de impacto.

ii. *métodos basados en la TRI:* las medidas del área entre las Curvas Características del ítem (CCI) estimadas en cada grupo (Linn *et al,* 1981; Mellenbergh, 1989; Raju, 1988, 1990; Rudner, 1977; Shepard, Camilli y Williams, 1984), la comparación de los parámetros de las CCI (Bock, Muraki y Pfeiffenberger, 1988; Hoijtink y Molenaar, 1992a; Kelderman, 1989; Lord, 1977, 1980; Muraki y Engelhard, 1989; Muthén y Lehman, 1985; Thissen y Mooney, 1989) y la comparación del ajuste de modelos (Linn y Harnisch, 1981; Thissen, Steinberg y Wainer, 1993).

Thissen, Steinberg y Wainer (1993) presentan una buena revisión de estos métodos, que gozan de muy buena reputación entre los psicómetras, si bien su complejidad computacional representa un coste importante a tener en cuenta antes de decidirse a su utilización.

iii. *métodos basados en tablas de contingencia:* métodos basados en χ^2(Camilli, 1979; Shepard, Camilli y Averill, 1981; Scheuneman, 1979), el procedimiento Mantel-Haenszel (Holland y Thayer, 1988), los modelos loglineales (Mellenbergh, 1982; van der Flier, Mellenbergh, Ader y Wijn, 1984), la regresión logística (Rogers y Swaminathan, 1993; Swaminathan y Rogers, 1990), la aproximación de la estandarización (Dorans, 1986; Dorans y Kulick, 1983, 1986; Dorans, Schmitt y Bleistein, 1992) y el procedimiento SBTEST (Shealy y Stout, 1993a, 1993b).

Dorans y Holland (1993) proporcionan una buena revisión de algunos de estos métodos. En castellano, se puede consultar el trabajo de Ato y López (1996) sobre el análisis estadístico de datos categóricos.

Al igual que los métodos basados en la TRI, son procedimientos muy reputados y, una buena parte de ellos, considerablemente más simples desde el punto de vista computacional, lo que los hace tremendamente atractivos. En particular, es de destacar por su sencillez, eficacia y por el peso que está cobrando en la literatura especializada el procedimiento Mantel-Haenszel.

iv. métodos basados en modelos de estructuras de covarianza (Camilli, 1992; Muthén, 1984, 1988), utilizándose básicamente el análisis factorial confirmatorio o restringido (Oort, 1992, 1993; Reise, Widaman y Push, 1993).

Estos métodos parecen ciertamente prometedores para la detección del FDI, aunque de momento no han recibido la atención prestada a los procedimientos anteriores.

Desde hace una década escasa, el tema del FDI puede considerarse como un tema estrella en la investigación psicométrica. Es raro encontrar en la actualidad un número de una revista especializada que no contenga al menos un artículo dedicado a este tema y, en particular, la cuestión más abordada en la literatura es la comparación de la eficacia de los distintos métodos en la detección del FDI. Simplemente, a título de ejemplo se pueden citar los trabajos de Barbero y Prieto (1997); Elosua, López y Torres (1997); Fidalgo y Mellenbergh (1995); Hambleton y Rogers (1989); Ironson y Subkoviak (1979); Kim y Cohen (1995); Raju, Drasgow y Slinde (1993); Rogers y Swaminathan (1993); Shepard, Camilli y Averill (1981) y Shepard, Camilli y Williams (1985), entre otros.

Otra cuestión que está recibiendo una atención creciente tiene que ver con un problema que surge siempre que existen ítems con FDI en un test: el problema de la contaminación de la medida de la habilidad. Muy brevemente: cuando un ítem presenta FDI, se está utilizando una medida contaminada de la habilidad para estudiar el funcionamiento diferencial de los ítems ya que la definición de FDI a partir de la que se opera es la de diferencias en la actuación en un ítem de sujetos pertenecientes a grupos distintos pero con la misma habilidad estimada. Se han propuesto distintos mecanismos de purificación de la habilidad con objeto de mejorar el proceso de detección de ítems con FDI para un buen número de métodos de detección. Así, Candell y Drasgow (1988); Candell y Hulin (1986); Hulin y Mayer (1986); Kim y Cohen (1992); Park (1988); Park y Lautenschlager (1990) y Segall (1983) para métodos basados en la TRI; Holland y Thayer (1988) para el procedimiento Mantel-Haenszel; van der Flier et al (1984) y Kok, Mellenbergh y van der Flier (1985) para los modelos loglineales y Gómez y Navas (1996) y Navas y Gómez (1994) para la regresión logística. Todos los estudios conducen a la misma conclusión: cuando se trabaja con una medida purificada de la habilidad, mejora sustancialmente la tasa de detecciones correctas e incorrectas en todos los métodos. Asimismo, comienzan a aparecer, todavía tímidamente, estudios en la literatura que tratan de investigar el posible funcionamiento diferencial de ítems con un formato distinto al tradicionalmente utilizado en este tipo de estudios: el formato dicotómico. En particular, se ha tratado el formato politómico (véase Gómez e Hidalgo, 1997; Hidalgo y Gómez, 1997; Hidalgo y López, 1997; Potenza y Dorans, 1995; Welch y Hoover, 1993), los ítems de respuesta construida (Dorans y Schmitt, 1993; Welch y Miller, 1995; Zwick, Donaghue y Grima, 1993) y también los tests adaptativos informatizados (Oshima, MeGenty y Flowers, 1994).

Un área mucho menos activa en la investigación sobre FDI es la relativa a sus causas, a la(s) explicación(es) más probable(s) del mismo, lo que nos devuelve nuevamente al tema del sesgo de los ítems. El primer escollo con el que nos encontramos es la desconcertante pero sistemática inconsistencia entre los ítems revelados como sesgados por el análisis estadístico (ítems con FDI) y los ítems que, según los expertos, pueden presentar problemas de sesgo. Berk

AFRIL

(1982) habla de diferencias irreconciliables entre los datos procedentes de análisis estadísticos y de análisis de juicio. Los tres lustros transcurridos no han mejorado ostensiblemente la situación. La realidad es que los expertos en minorías son incapaces de predecir por encima del nivel del azar qué ítems presentan dificultades para determinados grupos minoritarios y, además, apenas son capaces de explicar posibles fuentes de sesgo en ítems con índices FDI altos. Esto enfrenta al constructor del test con el siguiente dilema: "*¿el indicador estadístico revela un caso real de sesgo, evidenciando una laguna en la conceptualización del constructo del test o el alto índice de sesgo no es más que un artefacto estadístico, esto es, un signo inválido de sesgo?*" (Shepard, Camilli y Williams, 1984, p.96).

La formulación de este dilema nos pone sobre la pista de lo que pudiera ser la respuesta al problema de la explicación del sesgo: la conceptualización del sesgo en términos de una violación de la estructura dimensional propuesta para la prueba (Ackerman, 1988, 1992, 1993; Camilli, 1992; Kok, 1988; Linn et al, 1981; Oort, 1992; Oshima y Miller, 1992; Shealy y Stout, 1993a, 1993b) o, lo que es lo mismo, la conceptualización del sesgo en términos de validez. En efecto, "*la existencia de sesgo en una medida dada indica que no se acaba de comprender del todo el constructo que se está evaluando. Por tanto, los estudios de sesgo en la medida deberían de ser promovidos como parte del proceso general de la validez de constructo*" (Millsap y Ericson, 1993, p.329). Parece claro que un test válido dará lugar a medidas idénticas para sujetos con idénticos valores en la variable medida, con independencia de su grupo de adscripción. Por ello, el sesgo puede conceptualizarse como una fuente de invalidez que perjudica más a un grupo que a otro (Shepard, Camilli y Averill, 1981). Cole y Moss (1989) prefieren hablar del sesgo en términos de validez diferencial de una interpretación dada de las puntuaciones de un test para un determinado grupo de sujetos y, proponen un marco de referencia o trabajo que da cabida a una concepción más amplia de la información necesaria para Justificar la interpretación y el uso de las puntuaciones de los tests en los distintos grupos de interés. Esto es, proponen un marco que permite integrar las cuestiones técnicas relativas al sesgo o validez y las importantes cuestiones de política social igualitario, al incluir los valores como una dimensión más de las consideradas en el modelo.

Por su parte, Mellenbergh (1989) y Mellenbergh y Kok (1991) entienden que la búsqueda de explicaciones para el sesgo pasa por encontrar lo que el primer autor denomina rasgos sesgadores (*finding the biasing traits*) y proponen para esta búsqueda estrategias cualitativas, correlacionales, cuasiexperimentales y experimentales. En sintonía con este planteamiento se encuentra también el interesante trabajo de Oort (1992, 1993), formalizado en su teoría de los violadores potenciales (rasgos sesgadores de Mellenbergh): esta teoría proporciona un marco integrado que permite considerar de forma simultánea prácticamente la totalidad de los factores que pueden influir en la actuación del sujeto en el ítem y, consiguientemente, puede ayudar a entender porqué sujetos de la misma habilidad pueden tener distinta probabilidad de responder correctamente a un ítem que supuestamente mide esa habilidad. Estos violadores potenciales o rasgos sesgadores se pueden definir en términos de los contenidos y procesos mentales

AFRIL

implicados en el tipo de tarea implícita en el ítem y se relacionan fundamentalmente con la historia estimular de los sujetos, esto es, con el distinto bagaje cultural, educativo, social y económico de los sujetos pertenecientes a distintos grupos (Muñiz, 1990). En esta línea, Holland y Wainer (1993) proponen llevar a cabo estudios FDI en grupos mucho más acotados de sujetos que tengan en cuenta, por ejemplo, aspectos educativos y experienciales cuidadosamente definidos en la formación de los grupos de comparación.

Se puede convenir, por tanto, con Fidalgo (1996) en que *"esta visión global del sesgo de los ítems que tiene como base central la dimensionalidad de los tests psicológicos, constituye un marco correcto para afrontar el futuro bien pertrechado conceptualmente"* (p.445).

Otra línea de investigación apunta en la dirección de determinar aquellas características del ítem que se relacionan diferencialmente con distintos grupos de sujetos (véase O'Neill y McPeek, 1993 para una revisión reciente de la investigación sobre el tema). Sin embargo, hasta la fecha son pocos los factores que sistemáticamente muestras un patrón diferencial y, además, no está claro su mecanismo de actuación. Schmitt, Holland y Dorans (1993) proponen distintos procedimientos para evaluar hipótesis relacionadas con estos factores. De lo aquí expuesto parece deducirse de forma inequívoca que el FDI *"es un área fructífera de investigación sobre evaluación y uno de los pocos campos empíricos que nos puede decir algo nuevo sobre los tests, sobre qué miden y cómo lo hacen de bien"* (Holland y Wainer, 1993, p.422).

En resumen, en este tema, dedicado al análisis de la calidad de los ítems, hemos descrito el proceso a seguir, para depurar los ítems de un instrumento. Normalmente, este análisis de ítems se lleva a cabo en distintas muestras y una vez depurados los ítems, se pasan a una muestra distinta y se vuelven a analizar los resultados obtenidos. Cuando ya se suele admitir que están depurados se contrasta la fiabilidad y validez del instrumento y, si no es aceptable, se vuelve a empezar. En los dos temas siguientes describimos los aspectos más generales sobre la fiabilidad y validez de un instrumento de medida.

AFRIL