ÍNDICE

INTRODUCCIÓN

1. Teoría y Modelos 13
2. Concepto de Psicometría 14
3. La medición de los atributos psicológicos 17
   3.1. La psicología como ciencia cuantitativa 17
   3.2. Justificación del uso de los tests 18
   3.3. La definición de medición en Psicología ... 20
   3.4. La fundamentación de la medición psicológica 24
   3.5. La teoría de la medición conjunta 26

TEORÍA CLÁSICA DE LOS TESTS

1. El modelo lineal de Spearman 30
   1.1. El concepto de puntuación verdadera 30
   1.2. La variable aleatoria error 31
   1.3. Construcción del modelo clásico..... 32
2. Medidas replicadas, paralelas y tau-equivalentes 34
3. Relaciones basadas en medidas paralelas 36
   3.1. Criterios estadísticos para el paralelismo ... 37
   3.2. Conceptos fundamentales 38
4. El test como un compuesto 39
   4.1. Componentes paralelos: la fórmula de Spearman-Brown 40

4.2. Componentes tau-equivalentes: el coeficiente alfa 42
5. Procedimientos para la estimación de la fiabilidad 45
6. Factores que afectan al cálculo de la fiabilidad 50
6.1. Fórmula de cálculo: la fiabilidad como concepto unitario 50
6.2. Heterogeneidad del grupo 51
6.3. Tiempo límite 52
6.4. Longitud del test 53
7. Desatenuación..... 53
8. Estimación de parámetros e inferencias 55

EXTENSIONES DE LA TEORÍA CLÁSICA DE LOS TESTS

1. La Teoría de la Generalizalización de la aptitud . 57
1.1. Motivación de la TG y conceptos básicos . 58
1.2. Fuentes de error de medida y estimación de los componentes de varianza 60
1.3. Decisiones relativas y absolutas del Coeficiente G 63
1.4. Estudios de Generalización (G) y de Decisión (D) 65
2. El modelo factorial simple 67
2.1. Introducción 67
2.2. Condiciones para el establecimiento del modelo 68
2.3. La estimación de los parámetros del modelo . 69
2 4. El coeficiente omega 71
2.5. Ventajas de la aproximación del análisis factorial 73

LA TEORÍA DE RESPUESTA AL ÍTEM: MODELOS

1. Introducción 75
2. Conceptos implicados en la TRI 78
2.1. Dimensionalidad del espacio latente 79
2.2. Independencia local 81
2.3. No presencia de factores de velocidad 82

2.4. Curva característica del Ítem 83
2.5. Funciones de información 84
3. Los modelos dicotómicos unidimensionales: Modelos Básicos 85
3.1. Ojiva normal de Lord 85
3.2. Los modelos logísticos 87
3.3. El modelo de Rasch 89
4. Extensiones de los modelos básicos y nuevos enfoques 93
4.1. Modelos para ítems con formatos de respuesta politómicos 93
4.2. Modelos para el tiempo de respuesta 96
4.3. Modelos componenciales y multidimensionales 99
4.4. Modelos basados en suposiciones más flexibles 105
5. Un modelo unificado: GLIRT 110

ESTIMACIÓN DE PARÁMETROS EN LA TRI

1. Introducción 115
2. Procedimientos de estimación de la aptitud 116
2.1. Estimación por máxima verosimilitud (MLE) . 116
2.1.1. Propiedades 119
2.1.2. La función de información y la MLE . 120
2.2. Estimación Bayesiana 121
2.3. Procedimientos Numéricos 123
3. Estimación simultánea de los parámetros del ítem y de la aptitud 126
3.1. Estimación por máxima verosimilitud (MLE)... 128
3.1.1. Conjunta (JML) 128
3.1.2. Condicional (CML) 130
3.1.3. Marginal (MML) 131
3.2. Estimación Bayesiana 134
3.3. Otros métodos de estimación 136

PRINCIPIOS DE CONSTRUCCIÓN DE TESTS Y APLICACIONES

1. La construcción del test 141
2. Análisis formal de ítems 143
   2.1. Procedimientos basados en los modelos TCT . 143
   2.2. Procedimientos en la TRI 143
      2.2.1. Evaluación de los supuestos del modelo 144
      2.2.2. Evaluación de las características esperadas del modelo 147
      2.2.3. Evaluación de las predicciones del modelo 148
3. Selección de los ítems del test 150
4. Estudio de las características del test 151
5. Aplicaciones a tests específicos 152
      5.1. Construcción de bancos de ítems 152
      5.2. Estudio del sesgo de los ítems 153
      5.3. Igualación de tests 157
      5.4. Tests a la medida: Tests adaptativos informatizados 163

REFERENCIAS BIBLIOGRÁFICAS 169

PRINCIPIOS DE CONSTRUCCIÓN DE TESTS Y APLICACIONES

Los autores de esta obra creen que el conocimiento de los modelos psicométricos debe conducir al alumno a interesarse por cual es su papel en el proceso de construcción de los tests, así como su utilidad práctica. Es decir, cual es su función tanto en la estimación de las aptitudes de los sujetos como en el conocimiento del propio test como instrumento, así como su influencia en la detección y solución de los problemas que surgen en este tipo de estudios; el estudio de las diferencias individuales. Por lo tanto, introducimos unas breves pinceladas sobre estos dos campos, el de la construcción de los tests y el de sus aplicaciones.

1. LA CONSTRUCCIÓN DEL TEST

El proceso de construcción de un test requiere, en primer lugar, tener una definición concreta del rasgo que se quiere medir en los individuos. El tener definido el rasgo con precisión es una premisa ineludible, que debe ir acompañada por una clara exposición de cuales son las relaciones existentes entre ese rasgo y sus manifestaciones observables, al menos de aquellas a través de las que se va a realizar el proceso de la aproximación

a esa medida, ya que el rasgo se supone que no es directamente mensurable. En una segunda etapa, se procede a la enunciación de los ítems, en concordancia con lo establecido en la definición del rasgo y sus manifestaciones. La tarea de producción o invención de ítems suele estar a cargo de especialistas, que enuncian un gran número de ellos, sometiéndolos a un minucioso análisis crítico, o a un sistema de jueces. Se obtiene como resultado un grupo seleccionado de ítems, que sufren una nueva selección basada en procedimientos experimentales, es lo que denominaremos análisis formal de ítems. Una vez seleccionados los ítems que se consideran idóneos para la formación del test, se estudian las características del test resultante, así como el establecimiento de normas para la interpretación de las puntuaciones. En la TCT se hace siempre con referencia al denominado *grupo normativo* y en la TRI las inferencias se hacen en relación con la forma específica de las CCIs.

Esquemáticamente, en el desarrollo de un test se siguen las siguientes etapas (Santisteban, 1990):

- Elección del Rasgo y especificación de los objetivos del test
- Enunciación de un conjunto amplio de ítems
- Examen y juicio crítico pormenorizado de cada ítem
- Análisis formal de ítems
- Selección de los ítems del test
- Estudio de las características del test y estimación del rasgo
- Estandarización y/o establecimiento de normas para la interpretación
- El test como producto final

Las diferencias más importantes entre el desarrollo de tests utilizando la teoría clásica o la teoría de respuesta al ítem, se dan a partir del análisis formal de ítems, fase en la que se utilizan procedimientos muy variados dependiendo, tanto de la complejidad de los modelos considerados, como de las técnicas de análisis utilizadas. Veamos a continuación algunas fases del análisis, selección y estudio de las características del test.

2. ANÁLISIS FORMAL DE ÍTEMS

2.1. Procedimientos basados en los modelos TCT

Las técnicas de análisis clásico suponen la evaluación de la dificultad, del poder discriminante y de los distractores. El principal problema con la aproximación clásica es que los estadísticos de los ítems dependen en gran medida de las características de la muestra de examinados que se ha utilizado para el análisis. Las muestras heterogéneas, generalmente, permitirán obtener valores estimados más elevados del índice de discriminación, medido como coeficiente de correlación biserial o biserial-puntual. Las estimaciones de la dificultad varían en función de la aptitud media del grupo de examinados. Sin embargo, una gran ventaja de la TCT es que la estimación de los parámetros de los ítems es sencilla y requiere únicamente muestras de moderado tamaño para la obtención de estimaciones estables de los parámetros.

La detección de ítems *anómalos* (al menos para tests referidos a normas) se hace por procedimientos correlacionales. Un ítem inadecuado o anómalo es aquel que es demasiado fácil o demasiado difícil, o no discrimina, es decir, tiene una baja correlación ítem-test, en la población de examinados para la que se ha diseñado. Naturalmente, los valores estimados de los parámetros de los ítems cambiarán de una muestra a otra.

Otros índices que se utilizan en los análisis clásicos son los índices de fiabilidad y de validez, para la elección de aquellos ítems que maximizan la fiabilidad, o la consistencia interna, o la validez con respecto a un criterio (véase Santisteban, 1990).

2.2. Procedimientos en la TRI

Los procedimientos en la TRI son bastante diferentes a los que se emplean en la TCT. La principal ventaja de los métodos y procedimientos de la TRI es que, en teoría, permiten obtener parámetros de los ítems que son invariantes respecto a las mues-

tras utilizadas para su calibración. La mayor dificultad es que requieren muestras muy grandes para la obtención de estimaciones estables de los parámetros de los ítems, así como la complejidad matemática de las técnicas utilizadas en la estimación.

La eliminación de los ítems anómalos no es tan simple como en la TCT. En primer lugar, es necesario comprobar el ajuste del modelo elegido a los datos, descartando aquellos ítems que no presenten un buen ajuste. Para la determinación de cómo el modelo da cuenta de los resultados obtenidos en el test, deben tenerse en cuenta los siguientes aspectos (Hambleton y Swaminathan, 1985):

1) Evaluación de las suposiciones del modelo, ¿los datos del test satisfacen los supuestos del modelo?
2) Evaluación de las características esperadas, ¿se obtienen las ventajas que se esperan al utilizar el modelo de respuesta elegido?
3) Evaluación de las predicciones del modelo, ¿existe un buen ajuste entre las predicciones del modelo y los resultados observados en la realización del test?

2.2.1. Evaluación de los supuestos del modelo

La selección del modelo dependerá de que se cumplan los supuestos de los modelos. Si bien, aunque todos los modelos TRI adoptan supuestos fuertes, también muestran una gran robusted cuando se violan parcialmente algunas de ellas. Veamos a continuación como poner a prueba algunas de las suposiciones de los modelos básicos de la TRI.

a) Unidimensionalidad e Independencia local

El supuesto de unidimensionalidad del espacio latente es común a la mayoría de los tests construidos, ya que facilita la interpretación de las puntuaciones (Lumsden, 1976). Puesto

que la unidimensionalidad implica la independencia local (ver Hambleton y Swaminathan, 1985), basta con mostrar el supuesto de unidimensionalidad para aceptar que se cumplen ambos supuestos. Puede examinarse el supuesto de independencia local, evaluando la matriz de varianzas y covarianzas, o la matriz de correlaciones de las puntuaciones de los examinados, fijando diferentes intervalos para los distintos niveles de aptitud o de puntuaciones en el test (McDonald, 1981; Tucker, Humphreys, y Roznowski, 1986). Sólo cuando los valores de los elementos de la matriz fuera de la diagonal principal son pequeños, próximos a cero, se acepta el supuesto de independencia local.

Bejar (1980) propuso el siguiente procedimiento para evaluar la unidimensionalidad. En primer lugar, del test total se obtiene un subtest mediante consideraciones apriorísticas. Por ejemplo, basándonos en consideraciones de contenido, podemos aislar un subtest que parezca medir una habilidad diferente del resto de los ítems del test. En segundo lugar, se calculan los parámetros de esos ítems dos veces: una incluyendo todos los ítems del test en la calibración de los ítems y del test, y otra calibrando sólo los ítems del subtest. Finalmente, se comparan los dos conjuntos de parámetros estimados. En caso de que las estimaciones de los parámetros difieran se contradice la suposición de unidimensionalidad, puesto que la realización dependería del grupo en que está incluido el ítem.

Otro método es el propuesto por Horn (1965), en el que se comparan las representaciones gráficas de los autovalores de la matriz de correlaciones inter-ítem, con otra simulada, en la que los datos simulados son aleatorios, y se generan aproximando una distribución normal del mismo tamaño de muestra y el mismo número de variables que el test. En caso de unidimensionalidad, las dos representaciones han de ser similares, excepto en el primer autovalor de la matriz de datos reales, que será sustancialmente mayor (ver Drasgow y Lissak, 1983).

El análisis factorial también se utiliza para evaluar la suposición de unidimensionalidad para un conjunto de puntuaciones, utilizando correlaciones tetracóricas o phi. Las correlaciones te-

tracóricas tienen la característica de aportar una condición suficiente para la unidimensionalidad para un conjunto de ítems, que es que la matriz de intercorrelaciones tetracóricas se explique por un único factor común (Lord y Novick, 1968). Reckase (1979) propone representar los autovalores (del mayor al menor) de la matriz de correlaciones entre los ítems, para evaluar posteriormente, si existe o no, un factor dominante.

Stout (Nandakumar y Stout, 1993; Stout, 1987) propone el procedimiento estadístico DIMTEST para evaluar la unidimensionalidad esencial de datos binarios. Se entiende por unidimensionalidad esencial la existencia de un sólo factor dominante en las respuestas del test. La dimensionalidad esencial (d_E) de un conjunto de ítems *U* es la dimensionalidad mínima necesaria para satisfacer la asunción de independencia esencial. Según Stout un test (U_1, U_2, ...U_n) con longitud *n* se dice que es esencialmente unidimensional, si existe una variable latente □, tal que para los distintos valores de □:

$$\frac{1}{N(N-1)} \square_{1<i\pi j<N} \left[\operatorname{cov}(U_i, U_j) | \square \right] \square 0$$

El procedimiento DIMTEST proporciona una forma sencilla de evaluar la unidimensionalidad esencial de un conjunto de respuestas, construyendo estadísticos basados en las diferencias estandarizadas de varianzas calculadas en sus muestras para, finalmente, construir un estadístico que sigue una distribución normal, bajo la hipótesis de unidimensionalidad previamente establecida.

b) Igualdad en los índices de discriminación.

Esta suposición se hace sólo en el modelo de Rasch y se puede evaluar analizando la variabilidad de las correlaciones ítem-test, por ejemplo, mediante las correlaciones biserial puntual o biserial.

Otro procedimiento consiste en calcular el porcentaje de correlaciones ítem-test que caen fuera de cierto rango. Por ejem-

plo, Hambleton y Swaminathan (1985) proponen el rango igual a la media de las correlaciones ítem-test ± 0.15. Cuando dicho porcentaje está próximo al 100%, deciden que puede aceptarse la suposición de igualdad en los índices de discriminación.

c) No presencia de factores de velocidad

Hay que evitar la presencia de factores de velocidad que contaminen las estimaciones de la aptitud. Este supuesto se puede evaluar simplemente observando el número de examinados que no finalizan el test, aun cuando se suponga que los examinados han tenido suficiente tiempo para responder a todos los ítems. Hambleton, Swaminathan y Rogers (1991) proponen analizar los siguientes porcentajes, el de sujetos que terminan el test, el que completa el 75% del test y el número de ítems que han completado el 80% de los examinados. Existen procedimientos basados en indicadores más objetivos, para evaluar si un test es de velocidad o de fondo, que pueden consultarse en Santisteban (1990).

2.2.2. Evaluación de las características esperadas del modelo

La principal ventaja de los modelos de la TRI es la invarianza de las estimaciones de los parámetros:

a) *La invarianza de las estimaciones en el parámetro aptitud*, puede estudiarse mediante la administración a los examinados de dos (o más) conjuntos de ítems, de entre el conjunto total de ítems sobre los que está definida la aptitud, procurando que éstos tengan valores muy distintos en dificultad. Una vez que se ha estimado la aptitud para cada sujeto en los dos conjuntos de ítems, se representan gráficamente los pares de estimaciones de la aptitud. La representación gráfica debería ser una línea recta con pendiente unidad, puesto que la aptitud estimada para cada examinado, en teoría, no depende de los ítems particulares elegidos para la calibración (Wrigh, 1968).

b) *La invarianza de las estimaciones en los parámetros de los ítems* puede estudiarse comparando las estimaciones de esos parámetros, en dos o más subgrupos de la población de interés. Al igual que en el caso anterior, la representación gráfica de las estimaciones debería ser lineal (Shepard, Camilli, y Williams, 1984).

2.2.3. Evaluación de las predicciones del modelo

Al ser la aptitud una variable latente, las predicciones del modelo no se pueden comparar directamente con los datos observados, a excepción de los modelos de un parámetro. En estos modelos se estima sólo el parámetro dificultad, asumiendo que el poder discriminante de los ítems es idéntico para todos los ítems del test y que la conjetura es cero. Bajo estas suposiciones, el número de respuestas correctas es un estadístico suficiente para la aptitud, y las predicciones del modelo se pueden comparar directamente con los datos observados para cada grupo de puntuaciones. Existen distintos estadísticos implementados en distintos programas de ordenador para comprobar la bondad de ajuste del modelo de Rasch (Rogers y Hattie, 1987).

También se han construido estadísticos para comprobar la bondad de ajuste, para los modelos logísticos de dos y tres parámetros, que permiten comparar los datos observados con las predicciones del modelo. El procedimiento general para la construcción de tales medidas consiste en: (1) estimar la aptitud y los parámetros del ítem a partir de los datos (2) ordenar a los examinados de acuerdo con sus valores en aptitud, (3) formar subgrupos *ordenados* (4) calcular la proporción de examinados en cada subgrupo que responde correcta o incorrectamente a cada ítem, y (5) comparar estas proporciones *observadas* con las predichas por el modelo, utilizando un estadístico como chi-cuadrado y/o una representación gráfica. Una medida similar es la Q_1 de Yen (1981), que toma la forma:

$$Q_1 = \sum_{k=1}^{10} \frac{N_k\left(O_{jk} - E_{ik}\right)}{E_{ik}} + \sum_{k=1}^{10} \frac{N_k\left[\left(1 - O_{jk}\right) - \left(1 - E_{ik}\right)\right]^2}{1 - E_{ik}} =$$

$$= \sum_{k=1}^{10} \frac{N_k\left(O_{ik} - E_{ik}\right)}{E_{ik}\left(1 - E_{ik}\right)}$$

Se procede de manera que para cada ítem *i*, las proporciones observadas (O_{ik}) en cada intervalo *k* (*k*=1,..,10) en los que se divide la escala de la aptitud, el número de examinados sea similar. Posteriormente, las respuestas observadas para cada nivel de la aptitud se anotan en una tabla de contingencia 10 x 2. El valor esperado (E_{ik}) es la probabilidad media predicha de respuesta correcta en cada intervalo. Yen (1981) muestra que Q_1 se distribuye aproximadamente como chi-cuadrado con 10–m grados de libertad, siendo *m* el número de parámetros del ítem. Otra medida de ajuste similar es la de chi-cuadrado de Bock (1972), que difiere de la de Yen en que el número de intervalos no es fijo y en las proporciones esperadas que se computan utilizando la mediana (en vez de la media) de las estimaciones de la aptitud en cada intervalo.

McKinley y Mills (1985), construyeron un estadístico de razón de verosimilitudes G^2, que se distribuye igual que Q_1. Los examinados se ordenan por rangos y se dividen en 10 intervalos de igual tamaño, de acuerdo a las estimaciones de la aptitud, tomándose el número de respuestas correctas e incorrectas en cada intervalo para el cálculo de las O_{ik}.

$$G_i^2 = 2\sum_{k=1}^{10} N_k \left[O_{ik} \ln\left[\frac{O_{ik}}{E_{ik}}\right] + \left(1 - O_{ik}\right) \ln\left[\frac{1 - O_{ik}}{1 - E_{ik}}\right] \right]$$

También se han propuesto distintas representaciones gráficas para evaluar el ajuste del ítem, que se han utilizado conjuntamente con los estadísticos de ajuste, o como un diagnóstico exploratorio del ajuste del ítem. Hambleton y Swaminathan (1985) aconsejan una comparación gráfica de la media observada en el ítem con la predicha por el modelo, proponiendo que esta compara-

ción se represente junto a la función de respuesta al ítem, o a los residuos. Wainer y Mislevy (1990), proponen hacer representaciones similares, y Kingston y Dorans (1985) proponen el análisis de la regresión del rasgo sobre el ítem, como una herramienta de diagnóstico para el ajuste.

Orlando y Thissen (2000) presentan dos índices de la bondad de ajuste, que se basan en el método de Lord y Wingersky (1984), para predecir la distribución de verosimilitud conjunta en cada puntuación (número de aciertos). El procedimiento consiste en realizar una comparación directa de las frecuencias observadas y modeladas, tanto para las respuestas correctas, como para las incorrectas en cada nivel de puntuación (número de aciertos). Un algoritmo recursivo da la probabilidad conjunta para cada grupo y cada ítem.

3. SELECCIÓN DE LOS ÍTEMS DEL TEST

Cuando se aplican las técnicas clásicas para la construcción de tests referidos a normas, además de los aspectos concernientes a la validez de contenido, los ítems se seleccionan de acuerdo a dos características: la dificultad del ítem y su poder discriminante. En principio, se intenta elegir aquellos ítems con mayores índices de discriminación, ya que así la fiabilidad del test será mayor. La elección del nivel de dificultad se hace en función de los objetivos del test y de la aptitud supuesta en los examinados.

La TRI ofrece herramientas más poderosas para la selección de los ítems, realizándose de acuerdo a la cantidad de información que cada uno de los ítems proporciona al test. Lord (1977b) resume el proceso de selección mediante el uso de las funciones de información en los siguientes pasos:

1. Describir la forma de la función de información del test o *función de información objetivo*.
2. Seleccionar los ítems con funciones de información que mejor se ajusten al área bajo la curva de la información objetivo.

3. Calcular la función de información para los ítems seleccionados, cada vez que se añada un nuevo ítem.
4. Continuar con la selección de ítems, hasta que la función de información del test se aproxime de forma satisfactoria a la función objetivo.

Este procedimiento hace posible la construcción de tests con alto poder discriminante en una región particular del continuo de la aptitud. Cuando se tiene una idea de cual ha de ser la aptitud de los examinados, se pueden seleccionar los ítems que maximicen la información en la región en la que interesa hacer la evaluación, realizándose una selección óptima, con lo que se contribuirá sustancialmente a la precisión de la medida.

4. ESTUDIO DE LAS CARACTERÍSTICAS DEL TEST

En el estudio de las características del test, se incluyen los aspectos referentes a la validez del test y aquellos que hacen referencia a la precisión de la medida. Respecto a la validez (Alvarado y Santisteban, 2001) la visión actual tiende a ver la validez no como un análisis final de la calidad del test, sino como parte integrante y presente en todas las fases de la construcción del test. Por lo tanto, el estudio de la validez debe estar siempre presente, independientemente del modelo elegido para el desarrollo del test.

En cuanto a la estimación de la precisión de la medida, la TCT ha desarrollado distintos métodos para conocer la fiabilidad y el error típico de medida de los tests, en tanto que la TRI ha desarrollado el concepto de información para la medida de la exactitud de las estimaciones. Las ventajas sobre el concepto clásico son que su forma depende únicamente de los ítems incluidos en el test y no de la muestra concreta de examinados y que proporciona una estimación del error de medida para cada nivel de aptitud.

5. APLICACIONES A TESTS ESPECÍFICOS

5.1. Construcción de bancos de ítems

Un banco de ítems lo constituyen un gran conjunto de ellos, de los que se conocen sus propiedades psicométricas. Los ítems que forman parte del banco han de superar una fase previa de depuración y deben estar clasificados con respecto a algún criterio.

El disponer de un gran conjunto de ítems, no sólo facilita la tarea de construcción de tests y permite obtener tests de mayor calidad, sino que además supone un ahorro considerable en cuanto a medios y costes asociados.

La disponibilidad de programas de ordenador para el almacenamiento y recuperación de los ítems en una amplia variedad de formatos, así como para la realización de cambios en la composición del test de manera sencilla, ha supuesto un gran impulso en la elaboración de bancos de ítems, reflejado en monografías como la aparecida en la *revista Applied Psychological Measurement* (1996), o en la publicación de pautas para su uso correcto (por ej. Hambleton, Murray y Anderson, 1982; Ward y Murray-Ward, 1994).

Aunque los índices de dificultad y de discriminación del análisis clásico se han empleado para la selección de los ítems del banco, éstos tienen la desventaja de su limitada aplicabilidad, como consecuencia de su dependencia del grupo normativo utilizado en el proceso de calibración. Los parámetros que describen los items en los modelos de la TRI no tienen esta limitación y, en consecuencia, son muy superiores para el uso y descripción de los ítems que formarán parte del banco de ítems. El único inconveniente es que la media y la desviación típica de las puntuaciones de la escala de aptitud se establecen arbitrariamente, por lo que se hace necesario convertir todos los ítems a una métrica común, como se puede ver más adelante, bajo el título *igualación de tests* (epígrafe 5.3.).

Una vez que se dispone del banco de ítems las posibilidades para la elaboración de tests son enormes. Además del aho-

rro en costes de calibración, lo más importante es que se hace posible la construcción de tests con características especificas determinadas a priori. Por ejemplo, cuando se desea evaluar a sujetos de alta competencia, se elige como función de información objetivo la que dé la máxima información para los valores altos de la aptitud, incluyéndose en el test únicamente aquellos ítems con alto poder discriminante y alto nivel de dificultad.

Entre los programas más destacables para la elaboración y gestión de bancos de ítems reseñamos (ver Hontangas, 1999): *BANKIT* (Molina, 1997), *Interactive Assesor* (EQL International, 1996), *LXRTest* (Logic eXtension Resources, 1994), *MicroTest Professional* (Chariot Software Group, 1998), *Quiz Factory* (LearningWare, 1997), *Test Builder* (TopGrade Software, 1997), *Random Test Generator* (Hirtle, 1996) y *WinAsk Pro* (SmartLite Software, 1997).

5.2. Estudio del sesgo de los ítems

Al construir un test, además de asegurar su precisión y su validez, hay que asegurarse de que las puntuaciones están libres de sesgo. El sesgo ocurre cuando la ejecución en el test requiere fuentes de conocimiento diferentes de aquellas que se intentan medir, causando que las puntuaciones del test sean menos válidas para un grupo particular (Camilli y Shepard, 1994).

Aunque existen métodos para examinar el sesgo en los tests, éstos generalmente conducen al análisis de los ítems individuales y, para quienes construyen tests, el que los ítems individuales estén libres de sesgo, les indica su ausencia en el test. Por lo tanto, las investigaciones sobre sesgo típicamente se reducen a investigar el sesgo de los ítems. Ahora bien, para concluir que existe sesgo, no basta con encontrar evidencia de un menor rendimiento medio, por ejemplo, en un grupo minoritario, puesto que podrían existir diferencias legítimas en rendimiento, a consecuencia de una menor aptitud en ese grupo. Cuando ciertos subgrupos, o minorías, obtienen siempre puntuaciones

medias menores que los del resto de los examinados, no atribuibles a un menor nivel de aptitud, entonces se dice que el ítem está sesgado en contra de una minoría en particular, y a esta estrategia se le denomina «análisis del impacto adverso» (Dorans y Hollans, 1993).

Para distinguir entre evidencia empírica de un menor rendimiento de la conclusión de sesgo, se utiliza el término funcionamiento diferencial del ítem o *differential item functioning* (DIF). De modo que se dice que existe DIF cuando individuos pertenecientes a distintos grupos (subgrupo de interés o *focal* comparado con un subgrupo de referencia) pero con similar nivel de aptitud, no tienen la misma probabilidad de dar respuesta correcta al ítem (Lord, 1980).

Para la detección del DIF se han propuesto múltiples procedimientos (Berk, 1982; Millsap y Everson, 1993), que van desde las aproximaciones centradas en el ANOVA, hasta el empleo de técnicas derivadas de la TRI, como la $\square^2$ de Lord (1980), que contrasta la hipótesis nula de igualdad en los parámetros de los ítems en las poblaciones de referencia y focal; o procedimientos para computar la discrepancia midiendo el área entre las CCIs de ambos grupos (ver Santisteban, 1990). Desafortunadamente, los procedimientos para la detección del DIF que se basan en la TRI requieren muestras muy grandes, lo que no siempre es viable en la práctica, razón por la cual los investigadores se han centrado en el desarrollo de métodos para la identificación del DIF, que no exijan tamaños de muestra tan grandes y que a la vez sean eficientes y sencillos de implementar (Narayama y Swaminathan, 1996).

Entre los métodos más prometedores para la detección del DIF, Narayama y Swaminathan (1996) destacan los procedimientos no paramétricos de Mantel-Haenszel (MH) y el procedimiento de sesgo simultáneo del ítem (SIBTEST o SIB). El procedimiento MH y el SIB comparten muchas características, como su facilidad de implementación, el no requerir grandes muestras y proporcionar estadísticos con tests de significación. Por otra parte, cabe citar el procedimiento de regresión logística de Swaminathan y Rogers (1990), que tiene un funcionamiento similar

al MH y al SIB en la identificación del DIF uniforme y que además, tiene la ventaja de permitir la detección del DIF no uniforme.

Los términos DIF uniforme y no uniforme los acuña Mellenbergh (1982), que considera que el DIF es uniforme cuando no existe interacción entre el nivel de aptitud y la pertenencia a un grupo. Es decir, cuando la probabilidad de contestar correctamente a un determinado ítem es siempre superior en uno de los grupos, por lo que las CCIs nunca se cruzarían. Por otra parte, el DIF no es uniforme cuando existe interacción entre nivel de aptitud y grupo, es decir, la diferencia entre los grupos en la probabilidad de la respuesta correcta, no es la misma a lo largo del continuo aptitud, pudiéndose apreciar que las CCIs se cruzan. Rogers y Swaminathan (1993) distinguen a su vez entre DIF no uniforme y DIF no uniforme mixto. Este último término hace referencia a aquellos casos en que hay diferencia en los parámetros poder discriminante y dificultad, frente a los casos en los que sólo hay diferencia en el poder discriminante de las CCIs.

Veamos en que consisten algunos de estos procedimientos:

— **Mantel-Haenszel**. El método de Mantel-Haenszel (1959) se ha utilizado ampliamente en medicina y muy especialmente en análisis epidemiológicos y en otras ocasiones en que se emplean tablas estratificadas 2 x 2. Para la aplicación del método se comienza por una estratificación que permitirá eliminar o controlar de forma directa los sesgos en las comparaciones cuando existen variables de confusión (variable que enmascara o confunde la interpretación de los resultados de un estudio de asociación entre dos variables). La forma de proceder en el análisis es construir tablas 2 x 2 para cada uno de los estratos o subgrupos que se consideren. Se trata de «obtener estimaciones puntuales, estimaciones por intervalos de confianza y tests de hipótesis que resuman la información de los datos en un único valor numérico y dar una *odds ratio* de todos los datos estratificados» (Santisteban, 1999, p. 366).

En Psicometría fue introducido por Holland y Thayer (1988) para comparar las probabilidades de obtener una cierta respuesta correcta, entre los grupos focal y de referencia. Para re-

alizar la comparación se utilizan tablas de contingencia 2 X 2 en las que aparecen los datos de respuesta al ítem dadas por los miembros del grupo focal y el de referencia, construyéndose una tabla para cada ítem del test (grupo x ítem) en cada nivel de puntuaciones. En total han de construirse tantas tablas de contingencia como puntuaciones diferentes se hayan obtenido en el test. La hipótesis nula a contrastar es la de igualdad en las proporciones de sujetos que aciertan y fallan el ítem en cada una de las muestras, y para cada uno de los niveles en que se ha dividido la puntuación total del test. Este estadístico sigue una distribución chi cuadrado con un grado de libertad. La aplicación de este estadístico puede hacerse mediante el programa MHDIF (Fidalgo, 1994), que indica las diferencias entre los grupos mediante un índice. Recientemente, Gómez y Navas (2000), han encontrado evidencia de un mejor funcionamiento del procedimiento MH al compararlo con otros procedimientos basados en $\square^2$, como el modelo logit o la regresión logística y con técnicas basadas en la TRI (medidas de área). No obstante, el gran inconveniente de este procedimiento, ha sido su incapacidad para la detección del DIF no uniforme. Aunque recientemente se han publicado modificaciones a los procedimientos MH (Mazor, Clauser y Hambleton, 1994) y SIBTEST (Li y Stout, 1993), que permiten la detección del DIF no uniforme, los procedimientos más adecuados para la detección del DIF no uniforme son:

— **Logit iterativo**. Procedimiento en el que las respuestas a cada ítem se pueden predecir a partir de la pertenencia a uno de los grupos (focal o de referencia), del nivel de aptitud y de la interacción entre nivel de aptitud y el grupo de pertenencia. Es un procedimiento análogo al ANOVA en el que los niveles de aptitud se consideran como categorías discretas no ordenadas. Si el modelo que mejor ajusta a los datos, en ambos grupos, sólo incluye los efectos del nivel de aptitud mas una constante, se concluye que no existe DIF. Si además es necesario incluir el efecto del grupo, se trata de un caso de DIF uniforme. Si, por último, se debe incluir el efecto de la interacción nivel x grupo, se puede hablar de DIF no uniforme. Este procedi-

miento está implementado en el programa BIAS1 (van der Flier, Mellenbergh, Adèr y Wijn, 1984).

— **Regresión logística.** Swaminathan y Rogers (1990), advierten que el procedimiento de logit iterativo no tiene en cuenta la naturaleza ordinal de los niveles de aptitud, con la consiguiente pérdida de información, por lo que recomiendan el procedimiento de la regresión logística que viene expresado por la puntuación:

$$P(u = 1) = e^{z} / (1+e^{z}), \quad \text{siendo} \quad Z = \square_0 + \square_1\square + \square_2 g + \square_3(\square g)$$

donde □ es el nivel de aptitud observado de un individuo (por lo general, la puntuación total en el test), y *g* representa la pertenencia a uno de los grupos, tomando el valor uno para el grupo de referencia y el valor cero para el grupo focal, siendo □*g* el producto de las dos variables independientes □ y *g*.

El parámetro $\square_2$ corresponde a la diferencia entre los grupos en la realización del test, y el parámetro $\square_3$ corresponde a la interacción entre aptitud y grupo. Un ítem mostrará DIF uniforme sí $\square_2$ π 0 y $\square_3 = 0$, y DIF no uniforme sí $\square_3$ π 0, independientemente del valor de $\square_2$. Swaminathan y Rogers (1990), así mismo, proponen un estadístico para comprobar la hipótesis nula de la simultaneidad de $\square_2 = 0$ y $\square_3 = 0$. Este estadístico sigue una distribución $\square^2$ con dos grados de libertad.

El DIF compara la ejecución en el ítem de dos grupos, pero esa comparación debe hacerse sólo con los miembros con igual nivel de dificultad. Un problema importante es la determinación del DIF para ítems dicotómicos, pero es más complejo y extremadamente interesante el estudio del DIF para ítems politómicos con diferentes categorías de respuestas.

Además de los programas específicos que se han mencionado para la detección del DIF, este se puede analizar con paquetes estadísticos como el SPSS (regresión logística) o el SAS (regresión logística multinomial).

5.3. Igualación de tests

La comparación de las puntuaciones obtenidas en diferentes tests diseñados para la medida de una misma aptitud, es una cues-

tión de considerable importancia, tanto para los psicólogos que utilizan los tests, como para los expertos en medición psicológica. Además, una de las aplicaciones más importantes de la igualación de tests es la adaptación de tests de unos países a otros, proceso mucho más complejo que la mera traducción del texto.

La comparación de las puntuaciones obtenidas en distintos tests, se lleva a cabo mediante un proceso de *equiparación de puntuaciones*, de manera que puedan transformarse las puntuaciones a una métrica común. La necesidad de equiparación habitualmente surge de una de las dos situaciones siguientes:

1) La dificultad de los tests es comparable y las distribuciones de la aptitud de los examinados que han realizado los tests son similares.
2) La dificultad de los tests es diferente, así como la distribución de la aptitud de los examinados que han realizado las pruebas.

La equiparación, en el primer caso, se denomina *horizontal*, mientras que en el segundo caso se conoce como *vertical*. La equiparación horizontal es apropiada cuando se precisan varias formas de una prueba (por ejemplo, por motivos de seguridad), las distintas formas del test no tienen porqué ser *equivalentes* (Lord, 1980) pero se espera que sean paralelas. En contraste, la igualación vertical, es de interés para la construcción de una única escala que permita la comparación de las aptitudes de los examinados en diferentes niveles (por ejemplo, en distintos niveles educativos).

Lord (1980) introdujo el concepto *equidad*, para establecer las condiciones necesarias para que se dé la equivalencia entre los tests, es decir, que los tests puedan ser intercambiables, obteniéndose la misma aptitud estimada para los examinados. La equidad implica la no dependencia de la equiparación respecto a la muestra específica utilizada (invarianza), ni del test utilizado como base (simetría). Las condiciones de equidad son:

1) La equidad no puede darse en tests que midan diferentes rasgos o aptitudes.

2) La equidad no puede darse en tests con diferente fiabilidad.
3) La equidad no puede darse para las puntuaciones directas de tests con distinta dificultad (equiparación vertical), puesto que en este caso sus puntuaciones verdaderas no están linealmente relacionadas y, por tanto, los tests no tendrán la misma fiabilidad para los distintos niveles de aptitud.
4) La distribución condicionada de frecuencias de las puntuaciones x sobre □, $f[x|□]$, ha de ser igual a la de la puntuación transformada $x(y)$, o sea $f[x(y)|□]$, donde $x(y)$ es una función *uno-a-uno* de y.
5) En tests con fiabilidad distinta de uno, sólo podrá establecerse la equidad en el caso de que sean estrictamente paralelos.
6) La equidad puede establecerse para tests con fiabilidad perfecta.

En los estudios de equiparación se han seguido distintos diseños como son: (a) *un sólo grupo*, eligiéndose una muestra aleatoria de sujetos a las que se le aplican los dos tests a equiparar, (b) *dos grupos equivalentes*, en el que se eligen dos muestras aleatorias de la población, bajo la suposición de que el azar generó muestras equivalentes (la ventaja de este método es que elimina los efectos de la práctica y la fatiga), (c) *tests de anclaje*, aplicándose los dos tests a equiparar a dos grupos de sujetos. Esos tests deben contener un cierto número de ítems comunes *de anclaje*, que permitirá establecer las equivalencias entre los tests a equiparar. Los grupos no tienen por qué ser equivalentes, con lo que se evitan las dificultades de los otros métodos anteriormente mencionados. Como alternativa al diseño de anclaje, se pueden administrar dos tests a dos grupos de examinados, pero pasando ambos tests a unos cuantos individuos elegidos de ambos grupos.

Los métodos clásicos de equiparación los describieron en detalle Angoff (1971) y Kolen (1988). En general, estos modelos utilizan una de las dos equiparaciones siguientes: