2.- TEORÍA DE RESPUESTA AL ÍTEM

2.1.- INTRODUCCIÓN

Muchos conceptos, especialmente en el campo de las ciencias sociales y la medicina, no pueden ser observados de forma directa, ya sea porque se trata de un concepto abstracto o de una característica subyacente (**Calidad de Vida**, la inteligencia, la dependencia económica, etc.) o porque no cuenta con un valor exacto para su medición. Estos conceptos requieren la utilización de indicadores indirectos para su cuantificación, requieren ser analizados mediante técnicas estadísticas que definen una variable latente a partir de un conjunto de variables indicadoras (ítems). Entre estas herramientas, está la *Teoría de Respuesta al Ítem* (*TRI*).

La *TRI* como su propio nombre indica, analiza el comportamiento de los tests, no a un nivel agregado del test en conjunto, sino desagregado de cada ítem (SÁNCHEZ RIVERO, 2004). El rasgo latente (θ) o variable latente a medir se suele denominar "aptitud" ya que la misma proviene del ámbito de la educación en el que se trata de estimar la capacidad o aptitud de un sujeto para resolver las preguntas planteadas en un cuestionario. En el contexto de Calidad de Vida el rasgo latente (θ) o variable latente es el ***"Nivel de Calidad de Vida"*** del individuo necesario para seleccionar una categoría de respuesta de cada ítem del cuestionario. De esta forma, los individuos con bajos niveles de Calidad de Vida puntuaran las categorías más altas en el ítem.

En el contexto de la educación, las repuestas a cada ítem son "correcta" o "incorrecta", de forma que si se asocia a una variable para representar la respuesta del individuo, esta variable tomará el valor 1 si la respuesta es correcta, y 0 si la respuesta es incorrecta. En el contexto de la medicina, la "respuesta" a un ítem será presentar o no presentar un síntoma en un determinado grado, dado un nivel particular de Calidad de Vida del sujeto. De esta forma, los individuos con bajos niveles de Calidad de Vida tendrán probabilidades más altas de presentar un síntoma en un determinado grado.

A la probabilidad de contestar una categoría de respuesta de un ítem *i* a un determinado nivel de Calidad de Vida ($P_i(\theta)$), se le conoce con el nombre de función de respuesta al ítem o Curva Característica del Ítem (CCI).

La CCI muestra probabilidades cercanas a cero para valores pequeños del nivel de Calidad de Vida y probabilidades cercanas a 1 para valores elevados de la misma, produciéndose un incremento gradual de la probabilidad a medida que aumentan los valores del rasgo latente.

Además del parámetro θ, la probabilidad $P_i(\theta)$ va a depender también, en función del modelo que se proponga para su estimación, de otros dos parámetros, el parámetro de discriminación y el parámetro de dificultad.

El parámetro de dificultad b_i es aquel punto de la escala latente para el cual la probabilidad de responder correctamente al ítem es, del 50%. Nos indica la posición de la curva característica en la escala latente. Los valores de la escala latente se estandarizan; los valores del parámetro de dificultad, suelen variar de -2 a +2.

El parámetro de discriminación a_i está relacionado con la pendiente de la curva característica en el punto b_i, de manera que cuanto mayor sea la pendiente de la curva, mayores serán las diferencias en las probabilidades $P_i(\theta)$ de los valores latentes próximos.

La Figura 2.1 muestra una curva característica con sus parámetros fundamentales.

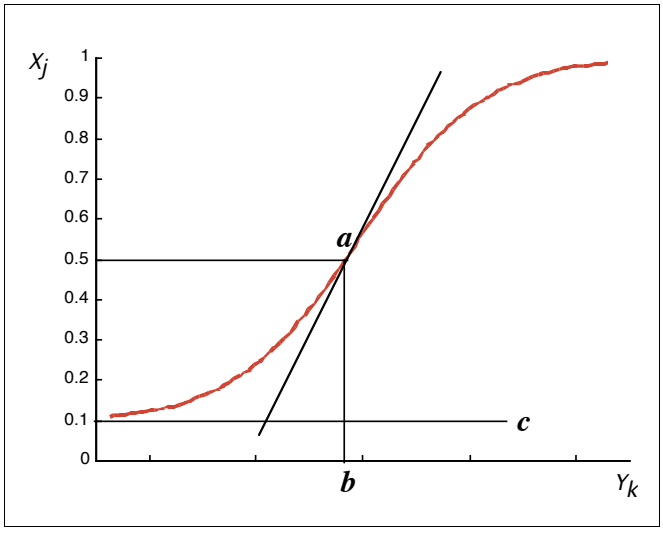

Figura 2.1.- Representación de los parámetros de la curva característica del ítem.

En términos de **Calidad de Vida** el parámetro de **discriminación** a_i se interpreta como la capacidad del ítem para discriminar o diferenciar entre distintos pacientes con distinta Calidad de Vida. El parámetro de **dificultad** b_i se interpreta como la puntuación en la escala de Calidad de Vida necesaria que debe de tener un paciente para que sea más probable que conteste una de las categorías del ítem. Los ítems con mayores parámetros de dificultad se corresponderán con aquellos que seleccionan los pacientes afirmativamente con Calidad de Vida más baja.

En algunas investigaciones, debido a que en la mayor parte de los tests, los ítems presentan respuestas categóricas, es un hecho que un ítem puede ser contestado por azar, por lo que se debe de tener en cuenta otro parámetros, el parámetro de adivinación c_i que se define como el valor $P(\theta)$ si θ tiende a $-\infty$. Gráficamente se corresponde con la asíntota inferior de la curva característica (Ver Figura 2.2).

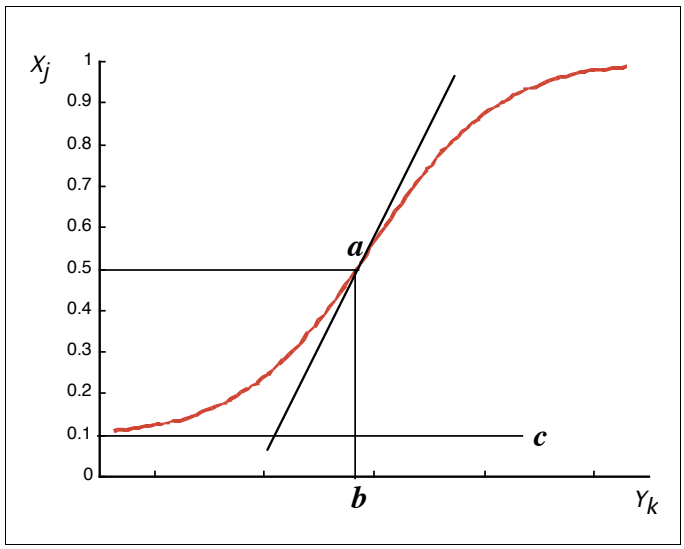

Figura 2.2.- Parámetro c en la curva característica del ítem.

Función de Información

La TRI utiliza el concepto de información del test, y de cada ítem, para reemplazar la fiabilidad. La información es una función que varía a lo largo de la escala y tiene forma de campana, si bien la función de información del test tiende a ser mucho más variable que la de los ítems particulares.

La Función de Información es una función definida para los valores de la variable medida (θ) que indica, para todos los posibles valores, con que precisión se está midiendo el test, o el ítem.

La **Función de Información del Ítem** indica que cantidad de información aporta el ítem a la medida de θ y a qué nivel aporta dicha información. La función de Información del Ítem en un nivel concreto (θ), es una función en dos aspectos: a) la discriminación del ítem, de modo que a mayor pendiente, mayor información; y b) el error típico del ítem en θ, de modo que a menor varianza mayor información (MARTINEZ, 1995).

La **Función de Información del Test** es un indicador de la precisión del test, ya que cuando mayor sea el valor del índice utilizado para su cálculo, menor será el error típico de medida y por consiguiente será mayor la información que las estimaciones aportan sobre el parámetro θ. BIRNBAUM (1968) definió la función de información de un test para un determinado valor de theta como la inversa de la varianza de los errores de medida para ese valor.

Los ítems con mayor poder para discriminar tienen informaciones más altas, aunque sobre un rango estrecho, mientras que ítems con menor poder para discriminar tienen menor información, aunque en un rango más amplio.

Supuestos de la TRI

Los modelos de Teoría de Respuesta a los Ítems constituyen un intento de superar algunos de los problemas con los que se enfrenta el enfoque clásico, y así mejorar la precisión de la medición. Para lograrlo, los modelos tienen que partir de supuestos más restrictivos, lo que a veces se denominan con el nombre genérico de teoría fuerte de los tests. Entre estas hipótesis hay dos que son fundamentales: la unidimensionalidad y la independencia local.

El supuesto de unidimensionalidad indica que la puntuación de un individuo en el test únicamente depende de una dimensión o factor: su nivel de habilidad en la variable medida. Esta es una asunción muy razonable, pues si estamos evaluando una

determinada variable, la medición obtenida sólo debería de depender del nivel de la persona en esa variable; pero constituye una restricción muy fuerte para muchas áreas psicológicas y educativas, en las cuales, los datos tienden a no ser estrictamente unidimensionales.

El estudio de la unidimensionalidad recibió un fuerte impulso en los años 80 por su importancia para los modelos más populares de la Teoría de Respuesta a los Items (TRI) (HAMBLETON y ROVINELLI, 1986).

El análisis factorial suele ser la metodología más utilizada para estudiar la dimensionalidad. Dado que empíricamente raras veces, si alguna, se encuentra una unidimensionalidad perfecta, esto es, que un solo factor dé cuenta de un 100 por 100 de la varianza, la unidimensionalidad se convierte en una cuestión de grado: cuanta más varianza explique el primer factor, más unidimensionalidad existe.

LUMSDEN (1961) propone como índice el cociente entre la varianza explicada por el primer factor y el segundo. Según LORD (1980) se necesitan procedimientos más rigurosos. HATTIE (1985) propuso 87 índices distintos basados en, modelos de respuesta (se basan en la idea de que un test perfectamente unidimensional es una función de la magnitud en que un conjunto de respuestas se desvian del patrón ideal), confiabilidad (el más utilizado es el coeficiente alpha Cronbach, 1951), componentes principales (el porcentaje de varianza explicada por la primera componente), análisis factorial (el valor del mayor valor propio) y basados en modelos de rasgos latentes.

En muchos casos los investigadores asignan valores enteros a cada categoría y proceden como si los datos fueran medidos en escalas de intervalo con las propiedades distribucionales deseadas, realizando una factorización de la matriz de correlaciones de Pearson para estudiar la dimensionalidad de los datos. De acuerdo con WAINER y THISSEN (1976) "una forma rápida y sencilla de proceder es suponer normalidad y estar en tu día de suerte".

Aunque muchos métodos estadísticos son bastante robustos frente a este tipo de desviaciones de las condiciones ideales, al menos en casos no demasiado extremos; hay situaciones en las que esta forma de proceder puede llevarnos a errores importantes. Por

ejemplo OLSSON (1979a), muestra que la aplicación de análisis factorial a datos discretos puede conducir a conclusiones erróneas relativas al número de factores y a estimadores sesgados de las cargas factoriales, especialmente cuando las variables observadas son asimétricas en direcciones opuestas. El problema se debe fundamentalmente a que los estimadores de las correlaciones son sesgados. Es necesario, por tanto, disponer de estimadores de la correlación cuando los datos son ordinales con solamente unos pasos de la escala OLSSON (1979b).

Se han llevado a cabo abundantes estudios de simulación para evaluar la robustez de los modelos TRI a violaciones del supuesto de unidimensionalidad (ANSLEY y FORSYTH, 1985; HARRISON, 1986; MUÑIZ y col., 1989).

Trabajos realizados por autores como MUÑIZ y CUESTA (1993) o CUESTA (1996) indican que los modelos de la TRI son bastantes resistentes a la violación del supuesto de unidimensionalidad.

La independencia local de los ítems es un concepto basado en el supuesto de que la respuesta a un ítem cualquiera; para un sujeto con un determinado nivel de habilidad, no afecta a las demás respuestas dadas a los otros ítems. En otras palabras, los ítems de un test que tienen el objeto de medir una variable unidimensional no pueden medir otra variable distinta. De acuerdo con esto podemos definir matemáticamente la independencia como el producto de las probabilidades de contestar a cada uno de los ítems que componen un test unidimensional, es decir:

$$P(x_1, x_2, ..., x_n / \theta) = P(x_1 / \theta) P(x_2 / \theta) ... P(x_n / \theta).$$

Si se cumple la unidimensionalidad, se deriva que existe independencia local entre los ítems; esto es, si se ha elegido la dimensionalidad correcta, los ítems son localmente independientes. Observamos por lo tanto que los conceptos de unidimensionalidad e independencia local son equivalentes (GOLDSTEIN, 1980).

2.2.- MODELOS DE LA TRI

Durante los últimos años, se han descrito y propuesto infinidad modelos para el uso en la Teoría de Respuesta al Ítem.

En un sentido muy amplio, los Modelos de Respuesta a los Ítems (MRI), son un conjunto de modelos, según los cuales las respuestas a los ítems de test dependen de una o más variables no observables continuas tal que, una vez el efecto de esas variables independientes observables es controlado, las respuestas a los ítems son independientes entre si (MAYDEU-OLIVARES, 1993).

Existe una amplia literatura sobre modelos unidimensionales de respuesta al ítem, es decir, con una única variable latente. Se han propuesto modelos unidimensionales para datos dicotómicos, politómicos ordenados (escalas Likert), y politómicos no ordenados; incluso, se habla de una taxonomía de los modelos de la TRI (THISSEN y STEINBERG, 1986).

En la mayoría de las investigaciones que usan modelos unidimensionales de la TRI utilizan dos tipos de funciones matemáticas para las CCI, la función logística y la curva normal acumulada. Así destacamos los modelos logísticos y los modelos de Ojiva Normal de uno, dos y tres parámetros.

2.2.1.- MODELOS LOGÍSTICOS

Modelo de Rasch o modelo logístico de un parámetro

Este modelo fue propuesto por el matemático danés RASCH (1960, 1966), utiliza la distribución logística, a partir de la aportación de BIRNBAUM (1968), para modelizar la probabilidad, en lugar de emplear la distribución normal, porque la primera es más manejable que la segunda, debido a su mayor simplicidad matemática.

La probabilidad de acertar un ítem *i* se puede modelizar utilizando la función logística de la siguiente forma:

$$P_i(\theta) = \frac{1}{1 + e^{-D(\theta - b_i)}}, \tag{2.1}$$

donde b_i es la dificultad del ítem *i*; D es una constante arbitraria; $P_i(\theta)$ es la probabilidad de acertar (responder) el ítem *i* a un determinada nivel de habilidad (a un determinado nivel de Calidad de Vida).

Modelo logístico de dos parámetros

El modelo logístico de dos parámetros fue propuesto por BIRNBAUM (1947, 1958a, 1958b, 1968). Este modelo añade para caracterizar los ítems otro parámetro, el parámetro de discriminación *a*. La ecuación es la siguiente:

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}}. \tag{2.2}$$

Modelo logístico de tres parámetros

El modelo logístico de tres parámetros fue propuesto por LORD (LORD y NOVICK, 1968; LORD, 1980), aunque tiene sus orígenes en los trabajos de BIRNBAUM (1947, 1958a, 1958b, 1968). La ecuación de este modelo es la siguiente:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}}, \tag{2.3}$$

donde c_i es un tercer parámetro que representa la asíntota más baja de la curva característica.

2.2.2.- MODELOS DE OJIVA NORMAL

Por modelos de ojiva normal entendemos aquellos en los que, en lugar de función logística, utilizamos la función de distribución de la normal estándar. RICHARDSON (1936) fue quien probó por primera vez el ajuste de ojiva normal a las respuestas de los ítems; posteriormente fue en los trabajos de LAWLEY (1943, 1944) donde se realizó

una descripción del modelo de ojiva normal como una función indicada para expresar la relación entre una dimensión latente tomada como continua con la probabilidad de respuesta correcta a un ítem; más tarde el trabajo de TUCKER (1946) también hace referencia al uso de la curva normal en las curvas características de los ítems. De la función de distribución normal acumulada se derivan los siguientes modelos:

Modelos de ojiva normal de uno, dos y tres parámetros

El modelo de ojiva normal de un parámetro se deduce del de dos parámetros al considerar que todos los ítems tienen el mismo poder discriminante, al parámetro a_i se le puede asignar un valor unidad, obteniéndose el siguiente modelo:

$$P_i(\theta) = \int_{-\infty}^{\theta - b_i} \frac{1}{\sqrt{2\pi}} e^{-0.5Z^2} dZ. \qquad (2.4)$$

El modelo de ojiva normal de dos parámetros es de la forma:

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-0.5Z^2} dZ. \qquad (2.5)$$

Por último, el modelo de ojiva normal de tres parámetros tiene por ecuación:

$$P_i(\theta) = c_i + (1 - c_i) \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-0.5Z^2} dZ. \qquad (2.6)$$

2.2.3.- MODELOS UNIDIMENSIONALES DE RESPUESTA POLITÓMICA

Otros modelos unidimensionales para ítems con respuesta politómica son: el Modelo de Respuesta Nominal (BOCK, 1972), el Modelo de Respuesta Graduada (SAMEJIMA, 1969), el Modelo de Crédito Parcial (MASTERS y WRIGHT, 1984) y el Modelo de Crédito Parcial Generalizado (MURAKI, 1992).

SAMEJIMA (1969, 1972, 1997) desarrolló una familia de modelos de respuesta graduada de clases latentes, los cuales pueden ser aplicados a categorías politómicas ordenadas. Estos modelos pueden ser utilizados para resolver preguntas planteadas en un cuestionario (por ejemplo, incorrecto, parcialmente correcto, correcto) o para medir actitudes y preferencias (por ejemplo una escala likert, muy buena, buena, satisfactoria, regular, mala, muy mala).

Esta familia de modelos incluye dos tipos de casos, los casos homogéneos y los casos heterogéneos. En el caso homogéneo, el poder discriminativo de los sujetos que contestan el cuestionario es constante a lo largo de todo el rango de latente (aptitudes, nivel de Calidad de Vida, etc.). En este caso las pendientes de las curvas características son idénticas y las funciones están posicionadas de forma ordenada con la puntuación del ítem. En el caso homogéneo se tiene la propiedad de la aditividad de las funciones de respuesta categórica.

En el caso heterogéneo, el poder discriminativo no es constante a lo largo del rasgo latente. En este caso las pendientes de las curvas características no son idénticas no se da la propiedad aditiva de las funciones de respuesta categórica.

2.2.3.1.- MODELO DE RESPUESTA GRADUADA

Entre los distintos modelos unidimensionales para datos politómicos ordenados, uno de los que más atención ha recibido hasta la fecha es el modelo de Respuesta Graduada de Samejima (1969). Este modelo es una extensión del modelo logístico de dos parámetros, para el caso en que la respuesta al ítem es politómica.

SAMEJIMA (1969) propuso los primeros modelos de respuesta graduada: el modelo de ojiva normal y el modelo logístico para datos de respuesta graduada (es decir, categorías politómicas ordenadas). Más tarde propuso un marco más amplio para modelos de respuesta graduada, distinguiendo el caso homogéneo, a los que pertenecen el normal ojiva y logístico y los casos heterogéneos (SAMEJIMA, 1972).

COHEN y col. en 1993, definieron "los ítems de respuesta graduada" como aquellos que tienen x_i respuestas categóricas ordenadas y permiten que un individuo elija una respuesta por ítem. Para este tipo de datos SAMEJIMA (1969) desarrollo funciones de probabilidad basadas en el modelo logístico de la TRI de dos parámetros.

El modelo se basa en las diferencias entre las funciones de respuestas categóricas, para un ítem con *m* respuestas categóricas, habrá *m-1* variables binarias, por ejemplo, para un ítem de cinco categorías, la primera variable binaria está entre individuos que seleccionaron una categoría frente a las cuatro categorías superiores, la segunda está entre individuos que seleccionaron la categoría 2 o una categoría más baja frente la categoría 3 o una categoría más alta y así sucesivamente, el proceso continua hasta construir cuatro categorías. No hay necesidad de calcular la quinta variable binaria porque el ítem será puntuado como un cero cuando un individuo no seleccione ninguna categoría.

Vamos a denotar, para un ítem *i*, la probabilidad de responder la categoría r o superior como $P_r^{i^*}(\theta)$, y para la categoría $r+1$ o superior como $P_{(r+1)}^{i^*}(\theta)$.

Para el modelo de respuesta graduada logístico (SAMEJIMA, 1969, 1972, 1997), la probabilidad para un ítem *i* de que un individuo responda a la categoría r o superior de un ítem *i* es:

$$P_r^{i^*}(\theta) = \frac{1}{1+e^{-Da_i(\theta - b_{(r-1)_i})}} \text{ para } r = 1,2,3,...,m \text{ y } P_1^{i^*}(\theta) = 1,\ P_{(m+1)}^{i^*}(\theta) = 0. \qquad (2.7)$$

La expresión de la *Función de respuesta categórica* para una respuesta determinada x, de un ítem *i* puede expresarse como:

$$P_r^i(\theta) = P_r^{i^*}(\theta) - P_{(r+1)}^{i^*}(\theta). \qquad (2.8)$$

Si se sustituye en la ecuación (2.8) la ecuación (2.7), la expresión de la *Función de respuesta categórica* de respuesta graduada para un modelo logístico viene dado por la expresión:

$$P_r^i(\theta) = \frac{1}{1+e^{-Da_i(\theta - b_{(r-1)_i})}} - \frac{1}{1+e^{-Da_i(\theta - b_{r_i})}}, \tag{2.9}$$

donde $P_r^i(\theta)$ es la probabilidad de que un individuo con un nivel de habilidad θ conteste a la categoría r de un ítem *i*; a_i es el parámetro de discriminación del ítem *i*; $b_{(r-1)_i}$ y b_{r_i} son los parámetros de dificultad para las categorías $r-1$ y r del ítem *i*; y *D* una constante.

El número de parámetros de dificultad (*b*) es uno menos que el número de categorías de respuesta (si un ítem tiene cinco respuestas categóricas, los parámetros de su curva característica serán también cinco, un único parámetro de discriminación y cuatro parámetros de dificultad).

Cada parámetro de dificultad, especifica la puntuación sobre la escala latente (θ) en la que el individuo tiene un 50% de posibilidades de responder una categoría de un determinado ítem o una categoría superior.

Para un ítem con cinco categorías de respuesta ordenadas, las diferencias en las probabilidades de respuestas, las cuales definen las curvas características para cada una de esas cinco categorías son:

$$P_1^i(\theta) = 1 - \frac{1}{1+e^{-Da_i(\theta - b_{1_i})}}, \tag{2.10}$$

$$P_2^i(\theta) = \frac{1}{1+e^{-Da_i(\theta - b_{1_i})}} - \frac{1}{1+e^{-Da_i(\theta - b_{2_i})}}, \tag{2.11}$$

$$P_3^i(\theta) = \frac{1}{1+e^{-Da_i(\theta - b_{2_i})}} - \frac{1}{1+e^{-Da_i(\theta - b_{3_i})}}, \tag{2.12}$$

$$P_4^i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_{3_i})}} - \frac{1}{1+e^{-Da_i(\theta-b_{4_i})}}, \tag{2.13}$$

$$P_5^i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_{4_i})}} - 0. \tag{2.14}$$

En la siguiente figura (Figura 2.3) vienen representadas las curvas características correspondientes a cada una de las categorías de respuesta del ítem *Durante los últimos siete días, ¿Cómo ha sido el dolor de espalda en el peor de los casos?* de la dimensión *Dolor* del cuestionario QUALEFFO que evalúa la Calidad de Vida en pacientes con Osteoporosis.

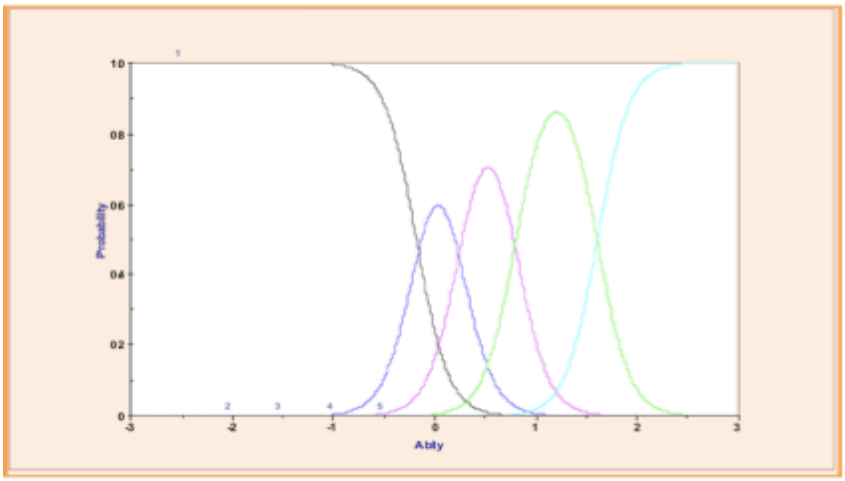

Figura 2.3.- Curvas características de un ítem de la Dimensión Dolor del cuestionario QUALEFFO

Los valores de los parámetros de discriminación y de dificultad son $a_3 = 6.17;\ b_{1_3} = -0.18;\ b_{2_3} = 0.25;\ b_{3_3} = 0.80;\ b_{4_3} = 1.61$. El gráfico anterior indica que para un individuo elegido al azar, la probabilidad más alta de responder a la primera categoría (*No he tenido dolor de espalda*) se obtiene para unos niveles de Calidad de Vida $\theta < -0.18$; la probabilidad más alta de responder a la segunda categoría (*Suave*) para unos niveles de Calida de Vida $-0.18 \le \theta \le 0.25$; a la tercera categoría (*Moderado*) para unos niveles de Calidad de Vida $0.25 \le \theta \le 0.80$; a la cuarta categoría (*Fuerte*) para unos niveles de Calidad de Vida $0.80 \le \theta \le 1.61$ y por último a la quinta categoría (*Insoportable*) para $\theta > 1.61$. La interpretación de estos parámetros se desarrollará de una forma más detallada en el capítulo cinco.

La Información del ítem es la inversa de la variabilidad del estimador máximo verosímil de θ en cada nivel y es aproximadamente $\frac{1}{s\varepsilon.(\theta)^2}$. Esta información aportada por el ítem es una medida local y para una categoría de respuesta *r* de un ítem particular *i* la *Función de información del ítem* para el modelo de Respuesta Graduada vale:

$$I_i(\theta) = \sum_{r_i}^{m_i} -\left[\frac{\partial^2 \log P_r^i(\theta)}{\partial \theta^2}\right] P_r^i(\theta). \qquad (2.15)$$

La *Función de Información del Test*, $I(\theta)$, se define como la suma de las funciones de información de los *n* ítems que componen el test

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta). \qquad (2.16)$$

En la siguiente figura (Figura 2.4) viene representada la Función de Información de una de las siete dimensiones que componen el cuestionario QUALEFFO, la dimensión del *Dolor*. El gráfico nos indica el nivel de información proporcionado por el test en cualquier nivel de aptitud. La máxima información para el cuestionario es 65.03, lo que significa que cada uno de los cinco ítems que componen esta dimensión debería de corresponder con una medida de información esperada de 2.03. El punto del rasgo latente en el que se alcanza la información máxima es aproximadamente 1. Además aparece el error estándar que da la misma información pero en términos de error estándar.

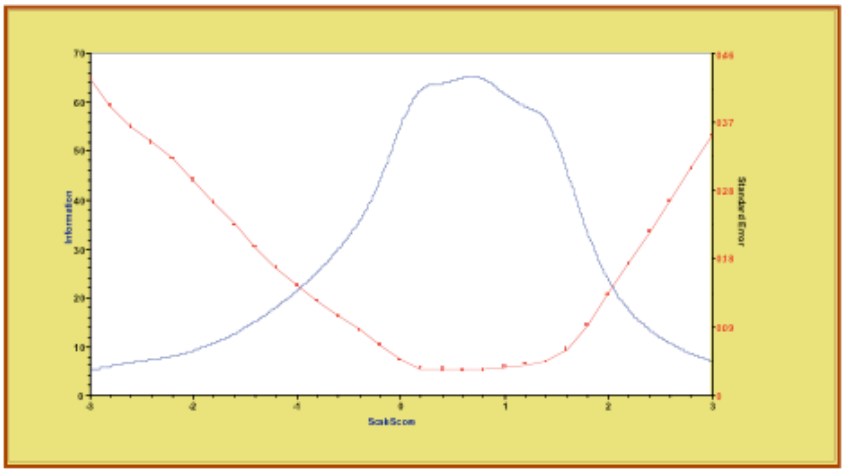

Figura 2.4.- Función de Información de la Dimensión del Dolor del cuestionario QUALEFFO

Con el modelo de Samejima, es posible calcular las puntuaciones esperadas para individuos sobre ítems o test de tipo Likert. La puntuación esperada para un individuo *j* sobre un ítem *i* se define como

$$t_{ji} = 1 + \sum_{r=1}^{m-1} P_i(\theta_j). \tag{2.17}$$

Tendríamos una expresión más complicada si los pesos para cada una de las categorías son diferente.

La puntuación esperada para un individuo *j* sobre todo el test se define como

$$T_j = \sum_{i=1}^{n} t_{ji}, \tag{2.18}$$

donde *n* es el número de ítems del test. Ambas funciones juegan un papel importante en el marco del *DFIT* que se desarrollará en el siguiente capítulo.

2.2.4.- ESTIMACIÓN DE LOS PARÁMETROS DE LOS MODELOS DE LA TRI

Seleccionado uno de los modelos, el paso siguiente será la estimación de los parámetros y de las puntuaciones θ de cada sujeto a partir de las puntuaciones empíricas en el test. Lo único conocido son las respuestas de los sujetos a los ítems del test y a partir de ellas realizarse la estimación de los parámetros. El proceso de estimación de los parámetros de los modelos de la TRI se denomina "Calibración".

En la actualidad se dispone de varios programas de ordenador a tal efecto, destacando entre ellos: BICAL, RASCAL, BILOG, MULTILOG. Todos estos programas ofrecen como salida fundamental los valores estimados de los parámetros de cada ítem y el valor de θ para cada sujeto.

El procedimiento de estimación de los parámetros más utilizado es el de *máxima verosimilitud* (BOCK y AITKIN, 1981; HARWELL y col., 1988), ya que como valores para los parámetros debemos de elegir aquéllos que maximicen la función de probabilidad de que ocurran los datos empíricos obtenidos; junto a éste se han utilizado procedimientos numéricos de aproximación como por ejemplo el de Newton-Raphson, el de " Scoring de Fisher", el algoritmo EM o diversos procedimientos de estimación bayesiana.

El Método de Máxima Verosimilitud se aplica a todos los tipos de modelos de la respuesta del ítem y es eficiente para las tests cortos y largos. La estimación se va haciendo por aproximaciones sucesivas (iteraciones). El proceso de iteraciones se detiene cuando los valores estimados de los parámetros convergen, es decir, cuando tras una iteración *n* no se producen cambios significativos en los valores estimados. Excepto en casos especiales, el método MML asume la independencia condicional para respuestas de diferentes ítems por individuos del mismo nivel de habilidad θ.

En la Teoría de Respuesta al Ítem, lo usual es considerar los estimadores Máximo Verosímiles Marginales (EML) que consisten en marginalizar la función de verosimilitud, integrando la función de densidad conjunta con respecto a los parámetros θ_j, obteniéndose las estimaciones máximo-verosímiles marginales a través de un proceso iterativo conocido en la literatura como algoritmo **EM**. Este algoritmo fue introducido explícitamente por HARTLEY (1958), en DEMPSTER, LAIR y RUBIN (1977) encontramos un extenso desarrollo de este método, razón por la cual en la literatura, se atribuye a estos últimos.

Se establece una clasificación de los métodos de máxima verosimilitud, según se consideren *condicionales* e *incondicionales*. Se denominan de alguna de estas formas dependiendo de que se obtengan estimadores condicionales o no de los parámetros. La mayoría de los autores en sus trabajos, utilizan para la estimación de los parámetros el método de máxima verosimilitud de los denominados incondicionales.

Los *métodos condicionales* son aquellos que usan la función de distribución de las puntuaciones dado un valor de la habilidad θ, o lo que es lo mismo, si se da el nivel de habilidad de cada uno de los individuos en la población, la función de verosimilitud está condicionada a este valores de θ_j. A través de estos métodos se obtienen estimadores condicionales de los parámetros, como lo hacen aquellos autores que siguen a ANDERSEN (1973).

Hay dos procedimientos bayesianos para estimar el nivel de habilidad (θ): la estimación máxima a posteriori (MAP) y la esperada a posteriori (EAP). Mientras que la estimación ML se fundamenta únicamente en los datos empíricos, los métodos bayesianos incorporan información sobre la distribución a priori de los niveles de habilidad de la población. La eficacia de cualquier método de estimación de parámetros de habilidad se valora a partir del sesgo y del error típico de medida que generan para diferentes niveles de habilidad y para distintas condiciones de aplicación.

2.2.4.1.- ESTIMACIÓN DE MÁXIMA VEROSIMILITUD MARGINAL PARA EL MODELO DE RESPUESTA GRADUADA

Recordemos que para el modelo de respuesta graduada logístico, la probabilidad de que un individuo *j* responda a una categoría *r* o superior de un ítem *i*, para un nivel de rasgo latente θ_j viene dada por la ecuación (2.7)

$$P_r^{i^*}(\theta_j) = \frac{1}{1+e^{-Da_i(\theta_j - b_{(r-1)_i})}}, \text{ para } r = 1,2,...,m$$

donde a_i es el parámetro de discriminación y $b_{(r-1)_i}$ es el parámetro de dificultad para la categoría $(r-1)$ del ítem *i*.

Además para este modelo se tiene que $P_1^{i^*}(\theta_j) = 1$ y $P_{(m+1)}^{i^*}(\theta_j) = 0$.

La función de respuesta, para una determinada categoría *r*, de un ítem *i* a un nivel del rasgo latente θ_j, viene dada por la ecuación (2.8)

$$P_r^i(\theta_j) = P_r^{i^*}(\theta_j) - P_{(r+1)}^{i^*}(\theta_j)$$

Consideremos para un individuo j el siguiente patrón de respuestas $\mathbf{x}_j = \left[x_{j1}, x_{j2}, ..., x_{jn}\right]$ donde para un ítem i con m respuestas categóricas, $x_{ji} = (x_{ji1}, x_{ji2}, ..., x_{jim})$ con $\sum_{r=1}^{m} x_{jir} = 1$.

Si asumimos independencia condicional, la probabilidad condicional de que un sujeto j responda el patrón $\mathbf{x}_j = \left[x_{j1}, x_{j2}, ..., x_{jn}\right]$ con un nivel de rasgo latente θ_j es

$$P(\mathbf{x} = \mathbf{x}_j / \theta_j) = \prod_{i=1}^{n} \left[P_1^i(\theta_j)\right]^{x_{ji1}} \left[P_2^i(\theta_j)\right]^{x_{ji2}} .. \left[P_m^i(\theta_j)\right]^{x_{jim}} . \qquad (2.19)$$

Para una muestra aleatoria con una función de distribución continua del rasgo latente θ, $g(\theta)$, la probabilidad incondicional viene dada por

$$P(\mathbf{x} = \mathbf{x}_j) = \int_{-\infty}^{\infty} P(\mathbf{x} = \mathbf{x}_j / \theta) g(\theta) d\theta . \qquad (2.20)$$

Esta probabilidad puede ser aproximada mediante una cuadratura de Gauss-Hermite por la suma:

$$\sum_{k=1}^{q} P(\mathbf{x} = \mathbf{x}_j / X_k) A(X_k), \qquad (2.21)$$

donde X_k es el punto de la cuadratura de la tabla de Gauss-Hermite (nodo) y $A(X_k)$ es el correspondiente peso (STROUD, SECHREST, 1966).

Consideremos ahora el siguiente patrón de puntuaciones observadas del ítem, para una muestra aleatoria de N individuos $l = 1,2,...,s$ donde $s \leq \min(N, m^n)$ y denotemos por v_l el número de individuos que responden al patrón anterior, entonces $\sum_{l=1}^{s} v_l = N$.

Ya que las frecuencias que asigna cada individuo a una y solo una de las m^n categorías del ítems, v_l siguen una distribución multinomial con N parámetros y $P_l = P(\mathbf{x} = \mathbf{x}_l)$.

El logaritmo de la probabilidad vale

$$\log L = C + \sum_{l=1}^{s} r_l \log P_l, \tag{2.22}$$

donde C no depende de los parámetros de los ítems.

Sustituyendo en la ecuación (2.22) las ecuaciones (2.19), (2.20) y (2.21), se tiene que

$$\log L = C + \sum_{l=1}^{s} v_l \log \sum_{k=1}^{q} \left[\prod_{i=1}^{n} \left[\left[P_1^i(\theta_j) \right]^{x_{ji1}} \left[P_2^i(\theta_j) \right]^{x_{ji2}} .. \left[P_m^i(\theta_j) \right]^{x_{jim}} \right] \right] A(X_k), \tag{2.23}$$

donde vamos a denotar por

$$\tilde{P}_l = \sum_{k=1}^{q} \left[\prod_{i=1}^{n} \left[\left[P_1^i(X_k) \right]^{x_{li1}} \left[P_2^i(X_k) \right]^{x_{li2}} .. \left[P_m^i(X_k) \right]^{x_{\lim}} \right] \right] A(X_k) = \sum_{k=1}^{q} L_l(X_k) A(X_k). \tag{2.24}$$

En la ecuación (2.24), $L_l(X_k)$ es la probabilidad condicional de $\mathbf{x}_l$ dado un valor del rasgo latente $\theta = X_k$.

Si derivamos la ecuación (2.23) respecto de los parámetros de un ítem *i*, $a_i, b_{1i}, ..., b_{ri}, ..., b_{(m-1)i}$, obtenemos un sistema de ecuaciones que resolveremos con el algoritmo EM.

Con el sistema resultante aparecen las ecuaciones de una regresión logística estándar, donde

1) X_k es el valor de la variable independiente al nivel *k*.

2) $\sum_{l=1}^{s} v_l x_{lir} L_l(X_k) A(X_k) / \tilde{P}_l = \bar{r}_{ikr}$ es la *frecuencia esperada* de responder a la categoría *r* del ítem *i* al nivel *k* dado x_{lir} y $P_r^i(X_k)$ (la suma de estas cantidades con respecto a *k* es el número de respuestas a la categoría *r* del ítem *j*).

3) $\sum_{l=1}^{s} v_l L_l(X_k) A(X_k) / \tilde{P}_l = \overline{N}_k$ es el *tamaño de muestra esperado* al nivel *k* dado $P_r^i(X_k)$ (la suma de esas cantidades con respecto a *k* vale *N*).

El algoritmo **EM** es un método que permite encontrar los estimadores máximo verosímiles de los parámetros de los ítems. El procedimiento numérico consta de dos pasos:

Paso E: para unos parámetros provisionales $b_{1i}, b_{2i}, ..., b_{(m-1)i}$ y a_i del ítem *i*, se calcula $L_l(X_k)$ con $k = 1,2,...,q$ y $\tilde{P}_l = \sum_{k=1}^{q} L_l(X_k)A(X_k)$ para el patrón l, $l = 1,2,...,s$. Se acumula $\bar{r}_{ikr}$ y $\bar{N}_k$ para calcular la suma con respecto a *l*.

Paso M: Se obtienen las estimaciones mejoradas de los parámetros $b_{1i}, b_{2i}, ..., b_{(m-1)i}$ y a_i, realizando un modelo logístico sobre $\bar{r}_{ikr}$ y $\bar{N}_k$ usando X_k como la variable independiente y la correspondiente ponderación.

Estos pasos del algoritmo **EM** se repiten hasta que las estimaciones sean estables, o hasta que los estimadores converjan. La convergencia sólo es geométrica, sin embargo, es lenta cuando se alcanza la solución. Para acelerar los cálculos se emplea un factor de aceleración propuesto por RAMSAY (1975) para la solución de las ecuaciones implícitas. Otro método consiste en acortar los vectores de respuesta dentro del mismo grupo de puntuación y calcular las probabilidades $L_l(X_k)$. Como muchas de las puntuaciones a los ítems son las mismas en los patrones dentro del mismo grupo de puntuación, el ahorro para el cálculo es considerable.

Los pasos del algoritmo **EM** también pueden derivarse como una extensión de del principio de *información faltante* utilizado por DEMPSTER, LAIRD y RUBIN (1977) para obtener una estimación de máxima verosimilitud cuando el modelo pertenece a la familia exponencial. En el contexto utilizado, la variable rasgo latente θ es la información faltante que, si fuese conocida, permitiría que los parámetros de los ítems fuesen estimados por el análisis convencional explicado anteriormente. Si el modelo pertenece a la familia exponencial, existiría un estadístico suficiente simple para θ, y, de acuerdo con el principio de información faltante, los valores esperados de este estadístico, dados los valores observados, serían sustituidos en el modelo logit.

En este caso no existe un estadístico suficiente simple para θ, pero si podemos reemplazar cada observación individual θ_j por la esperanza condicional, dada por la observación $\mathbf{x}_i$. Entonces, por el teorema de Bayes, la distribución condicional de θ dado $\mathbf{x} = \mathbf{x}_j$ es

$$g(\theta / x_j) = \frac{P(\mathbf{x} = \mathbf{x}_j / \theta) g(\theta)}{P(\mathbf{x} = \mathbf{x}_j)}, \tag{2.25}$$

y por lo tanto la esperanza condicional de θ dado $\mathbf{x} = \mathbf{x}_j$ es usando las ecuaciones (2.19) y (2.20)

$$E(\theta / \mathbf{x}_j) = \frac{\int_{-\infty}^{\infty} \theta g(\theta) \prod_{i=1}^{n} P_1^i(\theta) P_2^i(\theta) ... P_m^i(\theta) d\theta}{\int_{-\infty}^{\infty} g(\theta) \prod_{i=1}^{n} P_1^i(\theta) P_2^i(\theta) ... P_m^i(\theta) d\theta}. \tag{2.26}$$

Aproximando las integrales por la suma de *q*-puntos indexados por *k* como se hizo anteriormente, y recodificando el *j*-ésimo sujeto para el *l*-ésimo patrón, a partir de la ecuación (2.24) obtenemos

$$E(\theta / \mathbf{x}_l) \cong \frac{\sum_{k=1}^{q} X_k L_l(X_k) A(X_k)}{\tilde{P}_l}, \tag{2.27}$$

que es una simple media ponderada de la X_k. Hay *s* distintos patrones de respuesta, y por lo tanto *s* valores de $E(\theta / \mathbf{x}_l)$. El número de respuestas (el tamaño de muestra para un modelo) a las $\mathbf{x}_l$ es v_l, y el número de respuestas a la categoría *r* de un ítem *i*-ésimo es $\mathbf{x}_{lir} r_l$. Entonces el modelo puede ser ajustado para los *s* puntos, usando $E(\theta / \mathbf{x}_l)$ como la variable rasgo latente (esperada), con $\mathbf{x}_{lir} r_l$ el número de ítems a los que se responde a la categoría *r*. Como $E(\theta / \mathbf{x}_l)$ es una suma ponderada de los *q* términos en X_k, de los *s* valores para el rasgo latente solo nos quedamos con *q*.

El número de respuestas a la categoría *r* para el ítem *i*-ésimo para este valor del rasgo latente es entonces:

$\tilde{r}_{ikr} = \dfrac{\sum_{l=1}^{s} v_l \mathbf{x}_{lir} L_l(X_k) A(X_k)}{\tilde{P}_l}$ y el correspondiente tamaño de muestra $\overline{N}_k = \dfrac{\sum_{l=1}^{s} v_l L_l(X_k) A(X_k)}{\tilde{P}_l}$.

Para el modelo de respuesta graduada esto hay que hacerlo ítem a ítem, por lo que en el *E*-paso del algoritmo hay que sustituir el θ_j por su esperanza condicionada.

2.2.5.- BONDAD DEL AJUSTE DE LOS MODELOS DE LA TRI

No hay un acuerdo generalizado, entre los autores, en cuanto a la utilización de los estadísticos de ajuste; por lo que existe una gran variedad de procedimientos estadísticos para la comprobación del ajuste, si bien ninguno de ellos es totalmente satisfactorio. Ni siquiera todos los modelos admiten los procedimientos estadísticos habituales para el estudio de la bondad del ajuste, por lo que se proponen alternativamente procedimientos descriptivos para completar o sustituir a los procedimientos estadísticos.

En PARSCALE (MURAKI Y BOCK, 1993) se utiliza el estadístico de razón de verosimilitudes que siguen una chi-cuadrado como una medida de ajuste para cada ítem, y la suma de estos chi-cuadrados proporciona el estadístico chi-cuadrado para el test entero.

Notemos que el mejor ajuste de las curvas no sería usado como el único criterio en aceptar o rechazar un modelo específico; hay algunos modelos que están basados en principios absolutamente diferentes y todavía producen sistemas de curvas similares.

MULTILOG (THISSEN, 1991) computa las frecuencias esperadas para cada patrón de respuesta y entonces podemos se puede utilizar el estadístico likelihood-ratio para medir la bondad del ajuste.

La bondad del ajuste de un modelo de respuesta graduada se mide por el estadístico

$$G^2 = 2(\sum_{l=1}^{s} v_l \ln \frac{v_l}{N\tilde{P}_l}), \tag{2.28}$$

que sigue una distribución χ^2 con $s - mn$ grados de libertad (es decir, no se tienen en cuenta los patrones $r_l = 0$) para $mn < s < m^n$ o $s - mn - 1$ para $s = m^n$. Si m^n es grande en comparación con *N*, la tabla de frecuencias será escasa y (2.28) tiende a ser inestable porque $N\tilde{P}_l$ será inferior a 5 en la mayoría de los casos. En estos casos, las frecuencias de los patrones con menor esperanza deberían de agruparse hasta que todas las frecuencias esperadas serán superiores a 5. En principio, no importa que los patrones estén agrupados, pero a efectos de visualizar la frecuencia observada y la esperada, es conveniente poner en común las frecuencias contiguas cuando los patrones están ordenados de acuerdo a la estimación del rasgo latente que corresponde a cada uno. La agrupación debe de hacerse a un nivel mínimo de modo que el número de frecuencias, s_0, después de la agrupación sea mayor que mn. El test χ^2 con $s_0 - mn - 1$ grados de libertad después de agrupación, proporciona una prueba conservadora del ajuste del test ya que la probabilidad no ha sido maximizada en el conjunto de los datos.

2.2.6.- MODELOS MULTIDIMENSIONALES DE LA TRI

Los MRI multidimensionales, no asumen necesariamente que una única variable latente subyace a las respuestas observadas, sino que es necesario postular la existencia de varias variables latentes para representar adecuadamente los datos observables.

Se han propuesto modelos multidimensionales para datos dicotómicos, para datos politómicos ordenados y para datos politómicos no ordenados aunque para estos últimos no se ha implementado ningún procedimiento de estimación.

Los modelos multidimensionales se han clasificado en dos grupos, dependiendo a la manera como se define la dimensionalidad:

Modelos Compensatorios: son aquellos en los que si el sujeto posee un nivel alto en una de las dimensiones puede compensar un nivel bajo en algunas otras. Las dos funciones paramétricas, más utilizadas para este tipo de modelos son la cumulativa logística y la cumulativa normal, donde al contrario que en los modelos unidimensionales, la función normal, presenta claras ventajas sobre la función logística (MISLEVY, 1986).
Entre los modelos de este tipo está el de HATTIE (1981); el de DOODY-BOGAN y YEN (1983), y el propuesto por McKINLEY y RECKASE (1983).

Modelos No Compensatorios: son aquellos modelos en los que puntuar alto en uno de los rasgos no compensa los déficit en alguno de los otros, es decir, son aquellos que se utilizan cuando es necesario emplear varias habilidades simultáneamente para responder correctamente a un ítem. El modelo más popular ha sido propuesto por SYMPSON (1978).