Evaluar. Laboratorio de Evaluación Psicológica y Educativa. Facultad de Psicología
2004, n º 4 (setiembre) Universidad Nacional de Córdoba (Argentina). ISSN N º 1515 - 1867

SECCION METODOLOGICA

Teoría de Respuesta al Ítem: Supuestos Básicos

Cortada de Kohan, N.*

*Universidad del Salvador

Resumen: En este artículo se revisan los fundamentos del paradigma psicométrico de Respuesta al Ítem, lamentablemente poco empleado en nuestro medio, aunque posee ventajas considerables con relación al paradigma clásico, tales como las de generar medidas diferentes con ítems estrictamente comparables y no dependientes de las muestras específicas de investigación, así como alcanzar un verdadero nivel intervalar de medición. Recientemente ha aparecido en nuestro país y España un test de aptitud verbal construido bajo este paradigma, el Test Baires (Cortada de Kohan, 2003) por lo cual resulta apropiado incluir esta invitación al análisis de los postulados esenciales de la Teoría de Respuesta al Item.

Introducción

La Teoría de Respuesta al Item (TRI) (Rasch, 1963; Lord, 1980) intenta brindar una fundamentación probabilística al problema de medir constructos latentes (no observables) y considera al ítem como unidad básica de medición. La puntuación de una prueba en el modelo clásico estima el nivel de un atributo (aptitud, rasgo de personalidad, interés) como la sumatoria de respuestas a ítem individuales, mientras que la TRI utiliza el patrón de respuesta (Nunnally & Bernstein, 1995).

Recordemos que un test no es un instrumento de medición como un metro, un termómetro o un velocímetro que proporcionan mediciones directas en una escala numérica. Un test debe considerarse más bien como una serie de pequeños experimentos, en el que el examinador registra una serie de respuestas del examinado y estas respuestas no son mediciones directas sino que proporcionan los datos de los cuales se pueden inferir mediciones. Por lo tanto, como en cualquier experimento, existe el problema de controlar en las respuestas el error experimental. Este error

experimental en los tests surge de que en el experimento no operan solo las variables independientes sino otras conocidas como variables de error que pueden influir en las respuestas. Estas variables extrañas o de error deben ser controladas y hay para esto tres procedimientos fundamentales: 1) el apareamiento o estandarización, 2) la aleatorización y 3) el ajuste estadístico (Cortada de Kohan, 1998).

La TCT supone que las diferencias sistemáticas entre las respuestas de los examinados se deben solamente a la variación en la aptitud (es decir a las diferencias en su valor verdadero de la aptitud), y todas las otras fuentes potenciales de variabilidad debidas a los materiales o a las condiciones externas o internas de los examinados son mantenidas constantes por las técnicas de estandarización o bien tienen un efecto que es no sistemático (es decir aleatorio, al azar). Por consiguiente se controlan por el apareamiento o por a aleatoriedad, aunque esto implica una reducción de la validez externa. Las inferencias de los datos que producen los tests (como en cualquier experimento) no pueden ser generalizadas más allá de los niveles estandarizados de su error.

Los puntajes de dos tests diseñados para medir la misma aptitud, aunque se hayan estandarizado suelen ser distintos Esto se debe a que cada test tiene su propio conjunto de ítem y cada item tiene distintas propiedades. Desde el punto de vista de la medición, las propiedades de los ítem son variables de error que evaden la estandarización del test.

La limitación más importante de los tests elaborados según la teoría clásica es que no pueden separarse las características del examinado de las características del test: cada una puede ser interpretada solamente en el contexto de la otra.En la teoría clásica la aptitud se expresa por el puntaje verdadero. La aptitud de un examinado se define en términos de un test particular. Si el test es “difícil” el examinado aparecerá como de poca aptitud, si el test es “fácil” el examinado parecerá tener mucha aptitud. La dificultad de un ítem se define en este contexto como la proporción de examinados en un grupo determinado que contesta el ítem correctamente. Las características métricas del test (tales como confiabilidad y validez) se definen en términos de un grupo determinado de examinados con los que se ha construido el baremo o normas de interpretación de las puntuaciones. Esto implica que es muy dificultoso comparar examinados que tomaron distintos tests.

Para subsanar este problema los investigadores han usado el tercer método de control experimental es decir el del ajuste estadístico (Van der Linden & Hambleton, 1997) Este último requiere explícita parametrización de la aptitud que nos interesa así como de las propiedades de los ítem según un modelo que relacione sus valores con los datos de las respuestas recolectadas con el test. Si el modelo se sostiene y los parámetros de los ítem se conocen, el modelo ajusta los datos según las propiedades de los ítem del test y por lo tanto puede ser usado para producir mediciones de la aptitud que están libres de las propiedades de los ítem del test

Un test siempre se propone establecer inferencias sobre los rasgos psicológicos de los sujetos (no observables) basándose en la información que manifiestan en las respuestas. La teoría de la respuesta al ítem así como la teoría clásica de los tests consideran que cada individuo lleva asociado un parámetro individual, que en la teoría de la respuesta al ítem se denomina aptitud, incluyendo cualquier rasgo psicológico y se simboliza por la letra griega θ (zeta), y en la teoría clásica se denomina puntaje verdadero (V), que es inobservable pero que se puede estimar por sus manifestaciones observables que son los puntajes originales, X1, X2 ....Xn, de los tests.

La diferencia principal entre la teoría clásica de los tests (TCT) y los diversos modelos del rasgo latente o de teoría de la respuesta al item (en adelante la TRI) es que la relación entre el valor esperado y el rasgo en la TCT es de tipo lineal (X= V+e) mientras que en los diversos modelos de la TRI las relaciones pueden ser funciones de tipo exponencial, tales como los modelos de Poisson, de la ojiva normal, del error binomial, el modelo de Rasch o los modelos logísticos de 1,2 o 3 parámetros.

Conceptos fundamentales de la Teoría de Respuesta al Item

El origen de estos modelos puede encontrarse en Lazarfeld (1950), pero es en la década del 60 cuando la obra del danés George Rasch marca un hito en cuanto a generar investigaciones; aunque quienes han divulgado mejor estos modelos han sido Lord y Birnbaum en la obra de Lord y Novick (1968) y Lord (1980).

La TRI intenta dar una fundamentación probabilística al problema de la medición de constructos inobservables. Se considera acá el ítem como unidad básica

del test. Estos modelos son funciones matemáticas que relacionan las probabilidades de una respuesta particular a un ítem con la aptitud general del sujeto. Su origen no es tan nuevo: pero dada la complejidad de los cálculos para su aplicación solo empezó a difundirse y utilizarse gracias a programas de computación específicos como BIGSTEP, LOGIST, BILOG, entre otros.

La teoría clásica de los tests ha sido muy útil; todavía sigue teniendo validez pero se ha señalado que tiene dos grandes defectos. En primer lugar utiliza índices para los ítem cuyos valores dependen del grupo particular de examinados o muestra con los cuales fueron obtenidos y, en segundo lugar, las estimaciones de la aptitud o rasgo examinado (θ) dependen de la especial elección de los ítem seleccionados para el test (Nunnaly y Bernstein,1995).

En la teoría clásica se produce la paradoja de que un ítem es fácil o difícil según la aptitud de los examinados y la aptitud de los examinados depende de que los ítem del test sean fáciles o difíciles. Es decir que en la teoría clásica las características de los ítem son dependientes del grupo. Esto suele traer muchos problemas. Por ejemplo, consideremos dos examinados que contestan correctamente el 50% de las preguntas de dos tests que difieren en dificultad. ¿Pueden ser considerados igualmente capaces?, claramente, no. Además la teoría clásica está orientada hacia todo el test y por lo tanto no nos permite pronosticar cómo responderá un individuo a un ítem particular.
Los postulados básicos de la teoría de la respuesta al item o TRI son los siguientes:

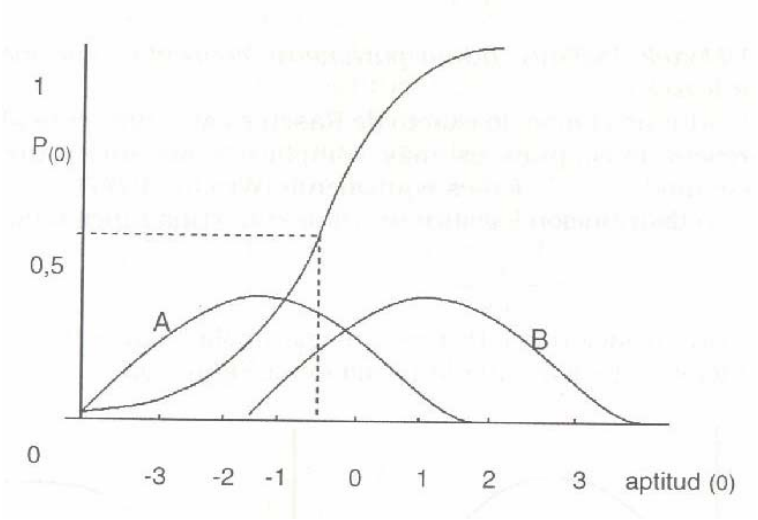
a. El resultado de un examinado en un ítem puede ser explicado por un conjunto de factores llamados rasgos latentes o aptitudes que se simbolizan por θ.

b. La relación entre la respuesta de un sujeto a un ítem y el rasgo latente que subyace puede describirse como una función monotónica creciente que se llama función característica del ítem o curva característica del ítem (CCI).Esta función especifica que a medida que la aptitud aumenta la probabilidad de una respuesta correcta al ítem también aumenta.

c. Las estimaciones de la aptitud (θ) obtenidas con distintos ítem serían iguales y las estimaciones de los parámetros de los ítem obtenidos en distintas muestras de

examinados serán iguales. Es decir que en la TRI los parámetros de aptitud y de los ítem son invariantes.Esta propiedad de invariancia se obtiene incorporando información sobre los ítem al proceso de estimación de la aptitud e incluyendo información sobre la aptitud de los examinados en el proceso de estimación de los parámetros de los ítem. La invariancia de los parámetros de los ítem puede verse claramente en la Figura 2.1, en donde se muestra la distribución de aptitud de dos grupos, A y B.

Figura 3.1. Distribución de aptitud en dos grupos poblacionales, A y B.



Los supuestos de la TRI son:

1. La unidimensionalidad del rasgo latente. Es decir que los ítem que constituyen un test deben medir sólo una aptitud o rasgo

2. La independencia. Es decir que las respuestas de un examinado a cualquier par de ítem son independientes y no existe relación entre las respuestas de un examinado a diferentes ítem . Así, las aptitudes especificadas en el modelo son los mismos factores que influyen sobre las respuestas a los ítem del test. De esta manera la probabilidad del tipo de respuesta a un conjunto de ítem es igual al producto de las probabilidades asociadas con las respuestas del examinado a los ítem individuales. Así, por ejemplo, si alguien tiene una probabilidad de 0,5 de responder correctamente cada uno de dos

ítem, la probabilidad de que el individuo responda correctamente a ambos ítem es de 0,5 x 0,5 = 0,25.

Existen muchos modelos de la TRI como ya comentamos anteriormente; pero acá nos limitaremos a los aspectos fundamentales de los que más se han difundido que son los logísticos.

1. Modelo logístico de un parámetro, más conocido como modelo de Rasch (1963. Aunque el modelo exacto de Rasch es algo diferente al que presentamos, pues es más complicado matemáticamente, este modelo logístico aquí propuesto es equivalente al original.

La distribución logística se define como una función tal que

$$y = \frac{e^x}{1 + e^x}$$

Su variación relativa es una parábola (Figura 3.2, a) y su función logística (Figura 3.2, b) es muy similar a la función normal acumulada. La CCI para el modelo de Rasch está dado por la ecuación siguiente:

$$P_{i\theta} = \frac{e^{(\theta - b)}}{1 + e^{(\theta - b)}} \qquad i = 1, 2, 3,....n$$

en donde:

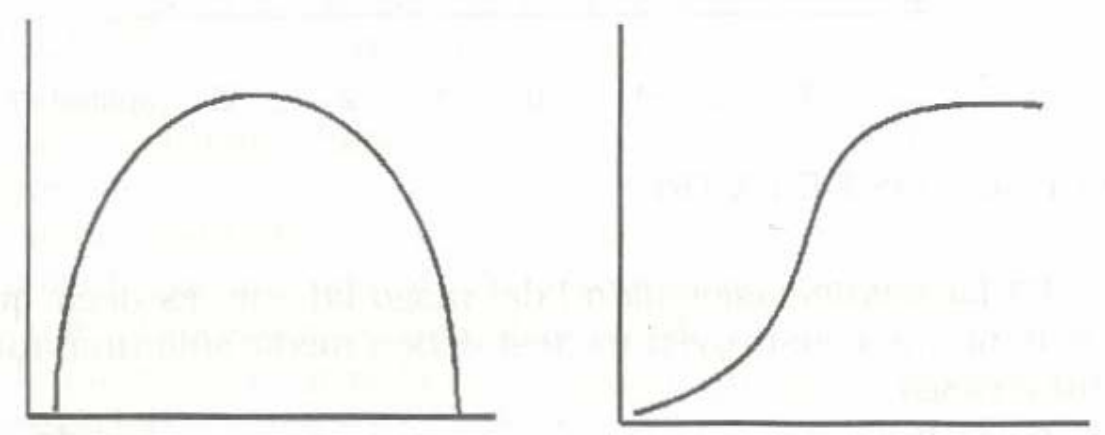
Pi(θ)= es la probabilidad de que un examinado elegido al azar con aptitud θ conteste correctamente el ítem i.

b = parámetro de la dificultad del ítem i

n = número de ítem del test

e = base de los logaritmos neperianos = 2,718

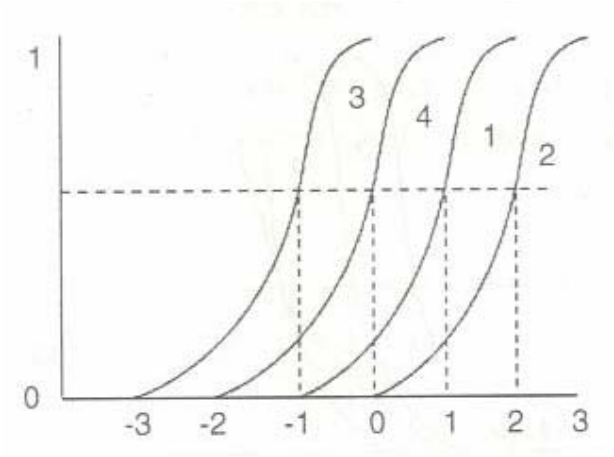
Figura 3.2 a) **Figura 3.2 b)**



La función forma una curva en forma de S con valores de 0 a 1 en la ordenada (probabilidad) y valores correspondientes a la aptitud θ en la abscisa.

El parámetro b de dificultad es el punto en la escala de aptitud θ cuya probabilidad de respuesta correcta es 0,5. Indica la posición del ítem en la escala de aptitud. Cuando más grande es el valor de b, mayor la aptitud requerida para que el examinado tenga una P = 0,5 de resolver correctamente el ítem. La aptitud suele transformarse de modo que la $\overline{X} = 0$ y la s = 1 y los valores de b suelen ir de -2 a + 2. Los ítem con b = -2 son muy fáciles , los ítem con b = +2 muy difíciles. En la Figura 3.3 presentamos el gráfico para 4 ítem tales que ítem 1 , b=1 ; item 2, b = 2 ; ítem 3, b= -1; ítem 4, b = 0.

Figura 3.3.



2. Modelo de dos parámetros.

Lord (1968,1980) fue el primero en elaborarlo, pero lo hizo basándose en una distribución normal. Actualmente este modelo es poco usado por su complicación matemática. En se sustituyó el modelo de dos parámetros de la ojiva normal por una función logística que tiene la ventaja de ser más conveniente para manejar. El modelo de la ojiva normal supone integración mientras que el modelo logístico no. Este modelo modificado está dado por la siguiente ecuación:

$$P_{i(\theta)} = \frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}} \qquad i = 1,2,3,...n$$

Aquí b es, igualmente que en el modelo anterior, el parámetro de posición o dificultad. El factor D = 1,7 es un valor arbitrario introducido para que la función logística sea ajustada a la ojiva normal con una exactitud de 0,01. Además hay un segundo parámetro a que es el de discriminación que es la pendiente de la CCI en el punto b. Los ítem con pendiente mayor son más útiles para separar a los examinados en distintos niveles de aptitud, que los ítem de menor pendiente. El modelo de dos parámetros es pues, una generalización del modelo de un parámetro. En la Figura 3.4 podemos ver las CCI de cuatro ítem con las siguientes características:

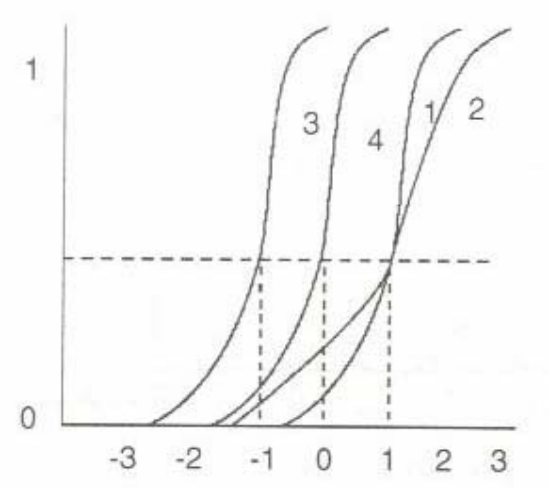
Item 1; b = 1 a = 1

Item 2; b = 1 a = 0,5

Item 3 ; b = -1 a = 1,5

Item 4; b = 0 a = 1,2

Figura 3.4.



Como puede apreciarse en el gráfico anterior las CCI no son paralelas como en el modelo de Rasch sino que en ciertos casos pueden cruzarse. Naturalmente los parámetros a y b pueden estimarse y tambien puede estimarse la P(θ) pero es un proceso algo complicado que solo puede realizarse con programas específicos de computación. Supongamos que tenemos un ítem para el que hemos obtenido los parámetros a y b y queremos saber la probabilidad en distintos puntos para trazar la curva CCI. El proceso en este caso sería el siguiente:

Item 55 a = 1,8 b = 1¿Cuál es la probabilidad del ítem en los valores de θ = -3, -2, -1 0, 1, 2, 3,?

Aplicando nuestros valores a la ecuación anterior, vale decir para θ = -3, tenemos:

$$P_{i(\theta)} = \frac{e^{1,7x1,8(3-1)}}{1+e^{1,7x1,8(3-1)}} = \frac{e^{6,12}}{1+e^{6,12}}$$

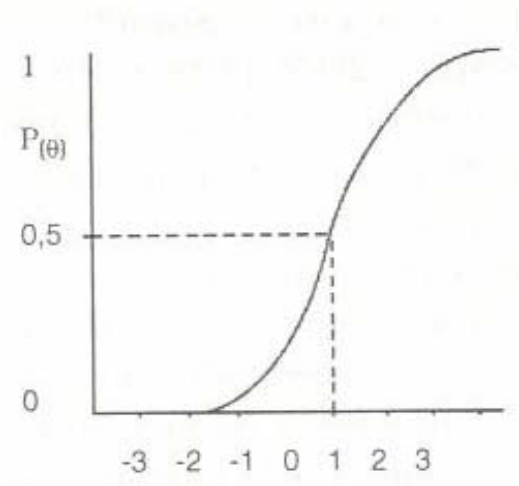
Usando los logaritmos neperianos (las calculadoras usuales los tienen)buscamos ex de 6,12 = 454,85 por lo tanto tenemos

P(θ) = 454,85 /455,85 = 0,9978

Repetimos esta operación para los distintos puntos de θ y podemos dibujar la curva característica del ítem (CCI) de la Figura 3.5 con los valores correspondientes a P(θ)

+3 = 0,9978
+ 2 = 0,9950
+ 1 = 0,500
0 = 0,0450
-1 = 0,0020
-2 = 0,0000
-3 = 0,0000

Figura 3.5



3) Modelo de tres parámetros.

La expresión matemática para este modelo es:

$$P(\theta) = c + (1-c)\frac{e^{Da(\theta - b)}}{1 + e^{Da(\theta - b)}} \qquad \text{i} = 1,2,3,...\text{n}$$

El parámetro nuevo acá es c que se llama parámetro del seudo azar porque representa la probabilidad en los ítem de opción múltiple de que un sujeto de poca aptitud conteste un ítem relativamente difícil de manera correcta, lo que hace suponer que lo hizo por azar, es decir adivinando. En este caso la curva es asíntótica en este caso.

Todos estos modelos sirven para aquellos tests en los que se puede considerar una respuesta correcta como 1 y la incorrecta como 0. Además de estos modelos existen otros modelos promisorios como el de las respuestas graduadas de Sameijim (1973) y el de Bock (1972) para escalas nominales.

La teoría de la respuesta al ítem se complica mucho para la estimación de los parámetros de los modelos. El proceso de estimación de los parámetros se denomina calibración. Es evidente que todos los procedimientos se hacen inmanejables sin la ayuda de los programas de computación. Actualmente existen varios programas tales como LOGIST y BICAL, entre otros. Por último es necesario puntualizar que para la emplear modelos TRI se requieren muestras grandes de sujetos (n>300) que hacen posible el ajuste a cualquier modelo de uno, dos o tres parámetros. Para muestras más pequeñas el mejor modelo es el de Rasch y por esto ha sido el más popularizado.

Una de las ventajas que más se ha señalado en la construcción de los tests de acuerdo a los modelos de la TRI es que se pueden elaborar tests individualizados es decir " a la medida" de los sujetos que permiten inferir en cada uno de los examinados un verdadero valor del rasgo de la manera más precisa.

Según el modelo del test logístico las curvas características de los ítem tienen la forma de una distribución logística acumulada:

$$Pi(\theta) = \left[1 + e^{-Dai(\theta - b)}\right]^{-1}$$

en donde (θ -b) se llama Li(θ) = 1,2,....n, o sea las unidades en una escala logarítmica o "logits".

ai = poder discriminante del ítem i

bi = nivel de dificultad del ítem i

Estos parámetros a y b son parámetros invariantes de un grupo de individuos a diferencia de lo que ocurre en el modelo clásico en donde dificultad y discriminación del ítem dependen de las características del grupo elegido. La relación entre las probabilidades de respuesta correcta e incorrecta al ítem viene dada por:

$$\frac{P_i(\theta)}{Q_i(\theta)} = e^{DL_{ii}}$$

tomando los logaritmos neperianos en esta expresión tenemos:

$$\ln \frac{Pi(\theta)}{Qi(\theta)} = DLi(\theta)$$

y esto representa una escala logarítmica en donde las unidades sobre la escala son "logits" (Recordemos que D = 1,7)

El método de estimación de los parámetros de los modelos de la TRI es el método de máxima verosimilitud que es un procedimiento que se hace con la ayuda de programas especiales de computación. En esencia si una variable X es una variable aleatoria continua con densidad f(x) la función de verosimilitud será :

$$L = f(x1)f(x2).....f(xn) = \pi f(x\theta k)$$

en donde L = (likelihood= verosimilitud) es la probabilidad de obtener la muestra observada y las f(x) son funciones de los parámetros θk. Los estimadores de máxima verosimilitud se obtienen resolviendo el sistema de ecuaciones:

$$\frac{\partial L}{\partial \theta_k} = 0 \qquad k = 1,2,3,...m$$

Se usan los logaritmos neperianos y se deriva respecto a θ y se iguala a 0. Es decir el principio de máxima verosimilitud es el de obtener estimadores de los parámetros desconocidos que maximicen la probabilidad de obtener las muestras.

Funciones de información y Funcionamiento diferencial del ítem.

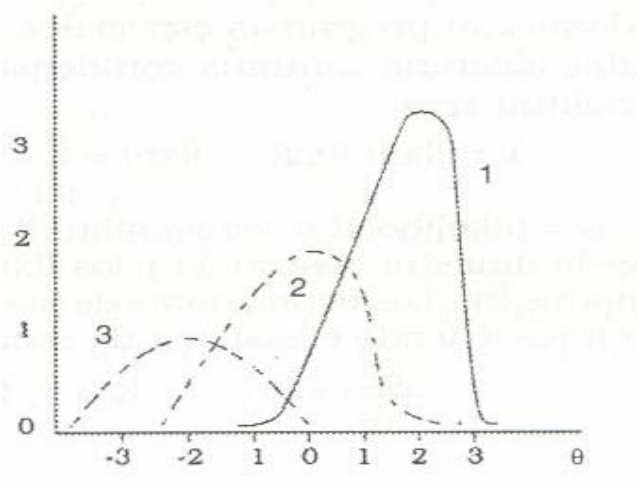
Una vez determinado el objetivo específico para el que se construye un test, el mejor test de k ítem a partir de n ítem disponibles es aquel que proporciona la mayor cantidad de información acerca del rasgo o aptitud latente.

La función de información de un test, I(θ : u i)se obtiene por la fórmula:

$$I(\theta iui) = \frac{Pi'^2(\theta)}{Pi(\theta)Q(\theta)}$$

siendo Pi(θ) la función de respuesta al ítem o CCI y P'i(θ) la primera derivada de Pi(θ) con respecto a θ. El valor de la función de información dependerá pues de la pendiente de la curva (cuando mayor más información) y del error estándar de medición (cuanto más pequeño mayor información) . Es decir, cuanto mayor sea la pendiente y menor la variancia de un ítem, mayor será la información. Esto se puede ver en la Figura 3.7 en donde tenemos el ítem 1 que da su mayor información en los niveles altos de aptitud, el ítem 2 en los niveles medios y el ítem 3 en los niveles bajos.

Figura 3.7.



La información para un determinado nivel de aptitud es directamente proporcional al poder discriminante del ítem. Generalmente los ítem dan su mayor nivel de información en los valores del rasgo latente próximos a su nivel de dificultad.

Los tests han sido algunas veces criticados por considerarse que eran sesgados respecto a las minorías étnicas y una de las ventajas de la teoría de la respuesta al ítem es que proporciona un marco de referencia unificado para conceptuar los sesgos a nivel de los ítem. Esto se consigue hallando lo que se denomina el funcionamiento diferencial del ítem o DIF. Un ítem presenta DIF si los sujetos que tienen la misma aptitud pero pertenecen a distintos grupos, no tienen la misma probabilidad de contestar bien el ítem. Por lo tanto el DIF puede investigarse comparando los parámetros que describen las curvas características del ítem. La hipótesis nula de que las funciones de respuesta del ítem son iguales se puede formalizar como:

$$H0;\ b1 = b2\ ;\ a1 = a2;\ c1 = c2$$

Donde 1 y 2 son los dos grupos que se comparan.

Para rechazar la H0 se necesitan estimaciones de los parámetros de los ítems y las matrices de variancia y covariancia. Luego se calcula la matriz de información para cada grupo y se invierten y se hace una prueba de $\chi 2$. Otra forma de obtener el DIF es

comparar las CCI directamente más bien que sus parámetros; si el área entre las dos CCI no es 0 quiere decir que existe DIF (Aguerri et al, 2002). Los métodos para detectar funcionamiento diferencial de ítem serán abordados en el capítulo de adaptación de tests.

Conclusiones

Sintetizando, Nunnally & Bernstein (1995) señalan tres ventajas fundamentales de TRI respecto a TCT:

- TRI permite comparar pruebas con reactivos diferentes, lo cual resulta el supuesto fundamental de las pruebas a medida o adaptativas
- Sujetos con un mismo puntaje en un test construido mediante el modelo clásico, en realidad difieren en su aptitud, problema que no se presenta en las pruebas TRI
- Las pruebas TCT y sus estimaciones del nivel de habilidad mediante el número de reactivos contestados correctamente no se relacionan linealmente con el verdadero nivel de aptitud del individuo. Por este motivo, una escala TCT no alcanza un nivel intervalar de medición.

Algunas voces críticas del modelo (Kline, 2000) enfatizan que, salvo para tests que evalúan dominios limitados (de rendimiento, por ejemplo), la teoría TRI no garantiza ítems unidimensionales. Otra dificultad de TRI es que requiere de muestras grandes (200 a 500 sujetos para escalas cortas) para calibrar los ítem. Además produce escalas muy cortas con alta homogeneidad que pueden ser inadecuadas para algunos propósitos de evaluación (Nunnally & Bernstein, 1995).

Una de las ventajas más destacables en los modelos de construcción de pruebas TRI es que permiten obtener tests personalizados, adaptativos o a la medida, a fin de inferir en cada uno de los examinados el verdadero valor del rasgo de manera más exacta. Otra de las aplicaciones útiles de la TRI es que se pueden construir bancos de ítem, vale decir un conjunto de ítem que miden una misma variable y cuyos parámetros están estimados en una misma escala (Attorressi et al, 1999). Estos ítem con sus parámetros se pueden almacenar y construir, de este modo, tests " a medida" según los objetivos de cada examinador. Por ejemplo, tests más difíciles para elegir becarios o tests más fáciles para ingresantes a una carrera, entre otras situaciones hipotéticas de evaluación. Una exposición autorizada y más

detallada de TRI, puede encontrarse en el texto de Hambleton & Swaminathan (1983): *Item Response Theory.*

Referencias

Aguerri, M. et al (2002). Evaluación de un método empírico para detectar el funcionamiento diferencial del ítem. *Interdisciplinaria*. *19* (2), 185-203.

Attorressi, H. et al (1999). Aplicaciones del modelo logístico de tres parámetros en una prueba de completar frases. *Investigaciones en Psicología*. *4*, (1), 7-25.

Bock, R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal cathegories. *Psychometrika*, 37, 29-51.

Cortada de Kohan, N. (1998). La teoría de la respuesta al ítem y su aplicación al test verbal Buenos Aires. *Interdisciplinaria*, *15*, 12, 101-129.

Hambleton, R. K. y Swaminathan, H. (1983). *Item response theory: Principles and applications.* Boston: Kluwer Academic Publishers.

Kline, P. (2000). *Handbook of Psychological Testing*. London: Routledge.

Lazarfeld, P. (1950). *The logical and mathematical foundations of latent structure analysis*. Princeton: Princeton University Press.

Lord, F. & Novick, M. (1968). *Statistical Theories of Mental Tests Scores*. New York: Addison Wesley

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale: Erlbaum Associates.

Nunnally, J. y Bernstein, I. (1995) *Teoría Psicométrica*. México: Mc Graw Hill.

RASCAL (1989). *Rasch item calibration program*. St. Paul: Assessment System Corporation.

Rasch, G. (1963). *Probabilistic models for some intelligence and attainment tests*. Copenhague: Denmarks Paedagogiske Institut.

Samejina, F. (1973). Homogeneous case of the continous response model. *Psychometrika*, 38, 203-219.

Van der Liden, W. & Hambleton, R. (1997). *Handbook of modern Item Response Theory*. New York: Springer.