Tema I: Modelos de Medición: Desarrollos actuales, supuestos, ventajas e inconvenientes.

Licenciatura de Psicología:
Desarrollos actuales de la medición: Aplicaciones en evaluación psicológica.
José Antonio Pérez Gil
Dpto. de Psicología Experimental.
Universidad de Sevilla.

Tema 1

Teoría de Respuesta a los Items (TRI).

1.1. Introducción.

1.2. Supuestos fundamentales.
- 1.2.1. Unidimensionalidad del espacio latente.
- 1.2.2. Independencia local.
- 1.2.3. Otros supuestos.
- 1.2.4. Curva Característica del ítem.

1.3. Modelos Básicos.
- 1.3.1. La contribución de Lord: el modelo de ojiva Normal.
- 1.3.2. La contribución de Rasch: el modelo logistico de Rasch.
- 1.3.3. La contribución de Birnbaum: los modelos logísticos.

1.4. Desarrollo de los modelos básicos y su clasificación.

1.5. Comprobación de los modelos.
- 1.5.1. Estimación de parámetros.
- 1.5.2. Ajuste del modelo.

1.6. Curva Característica del Tests.
- 1.6.1. Puntuación verdadera en el test.
- 1.6.2. Error típico de medida.

1.7. Función de información.

1.8. Ventajas y limitaciones.

1.1. Introducción.

Entre los desarrollos actuales de la medición, la Teoría de Respuesta a los Items (TRI) es, sin lugar a dudas, una verdadera alternativa que ha proporcionado las mejores soluciones a los problemas que presenta el modelo clásico, y representa un cambio real de modelo en el campo de la teoría de los tests; de hecho, como señala Muñiz (1997), esta alternativa supone un cambio radical a los planteamientos de la TCT, aunque no llega a ser una teoría contrapuesta sino complementaria al modelo clásico, es decir, la TRI ofrece soluciones a la mayoría de los problemas que tienen planteado la TCT desde hace años, da mejores respuestas a los propios planteamientos de la TCT y presenta aspectos que son imposibles de plantear desde la TCT. En este sentido, de acuerdo con Navas (1997), puede hablarse de un posible relevo paradigmático en el campo de la Teoría de los Tests. Esta teoría es la que domina en la actualidad y representa el mayor avance en la medición psicológica y educativa en los últimos años (Muñiz y Hambleton, 1992).

La TRI surgió en los años cincuenta como una reacción a los problemas y limitaciones que presentaba la TCT (Lord, 1952). No obstante, sus antecedentes podemos ubicarlos en la década de los años veinte, en concreto, en los trabajos de Thurstone (1925,1927,1928), donde aparecen los primeros esbozos de lo que será el concepto de curva característica del ítem. Thurstone (1928), afirmaba que *"el objeto al que se aplica la medición no debe influir mucho sobre la función propia del instrumento de medición. ...Dentro de la gama de objetos susceptibles de ser medidos mediante ese instrumento, la función de éste debe ser independiente del objeto sometido a medición"* (Thurstone, 1928, recopilado en Wainerman, 1976, p. 283). Esta idea fue tomando cuerpo paulatinamente durante las décadas de los años treinta y cuarenta en las que aparecen trabajos esporádicos con aportaciones como las de Richardson (1936), Ferguson (1942), Lawley (1943, 1944), Brodgen (1946) y Tucker (1946).

Es en la década de los cincuenta cuando tiene lugar el surgimiento de esta teoría con la publicación, en un número monográfico de la revista *Psicometrika*, de la tesis doctoral de Lord: *A theory of test scores* (1952)*;* en ella presenta formalmente las bases teóricas del modelo de ojiva normal de dos parámetros. A esta aportación se le une la introducción del modelo de la distancia latente y del modelo lineal, formulados por Lazarsfeld, (1950, 1959), que no han tenido tanta transcendencia pero que contribuyeron a clarificar y ampliar el marco teórico de esta nueva teoría. En la década siguiente, los sesenta, tienen lugar las publicaciones de los trabajos de Rasch (1960) y de Birnbaum (1968) donde se desarrollan los modelos logísticos. Así, Rasch presenta el modelo logístico de un parámetro, en su obra *Probabilistic models for some intelligence and attainment tests*, y, Birnbaum presenta, en el libro *Statistical theories of mental test scores* de Lord y Novick, los modelos logísticos de dos y tres parámetros. Estos modelos, van a posibilitar un tratamiento matemático asequible y facilitarán el desarrollo de métodos de estimación de parámetros. Los trabajos de Lord, Rasch y Birnbaum pueden considerarse como la puesta en escena de la TRI en el mundo académico. Sin embargo, habrá que esperar todavía algún tiempo para ver sus aplicaciones prácticas; efectivamente, a partir de sus aportaciones se producen una proliferación de trabajos desde la óptica de estos modelos; se amplía cada vez más su campo de aplicación y, en las revistas más importantes del área, aparecen monografías sobre el tema; se desarrollan diversos métodos de estimación de parámetros (Wrigth y Panchapakesan, 1969; Lord, 1974, Bock, 1972) y nuevos modelos para distintos formatos de respuestas, entre los que destacan el modelo de respuesta graduada (Samejima, 1969), el modelo de respuesta continua (Samejima, 1972) o el modelo de respuesta nominal

Bock (1972). La implantación y transición definitiva de la TCT a la TRI tiene lugar en la década de los años ochenta, cuando Lord publica su obra *Applications of ítem response theory to practical testing problems*, y, desde entonces domina en Psicometría (Baker, 1989) y se instaura como el marco de referencia desde el que se abordan los problemas de la medida psicológica y sus aplicaciones abastecen la mayoría de las necesidades prácticas de la medición.

Las causas que pueden explicar que haya tenido que esperar tanto tiempo para lograrlo, a pesar de que sus orígenes son solo un poco posteriores a los del modelo clásico, se deben, como señalan Navas (1997) y Hontaga (1997), por un lado, a que la TRI no se desarrolló en un contexto vinculado a las teorías de la inteligencia (como ocurrió con la TCT) sino vinculado a problemas técnicos en la construcción de tests y en la estadística matemática (Embretson, 1985) y, por otro lado, a su complejidad matemática, puesto que no llega fácilmente a lectores no iniciados, pero, sobre todo, a que el soporte matemático, informático y tecnológico que necesita esta teoría ha hecho que carezca de procedimientos prácticos necesarios para su implementación (Jaeger, 1987). Gracias al avance espectacular de la informática se posibilitó implementar de manera eficiente los métodos de estimación de parámetros ya desarrollados y, en esta epoca, aparecieron las primeras versiones de los programas BICAL (Wright y Mead, 1976), ANCELLES, OJIVA (Urry, 1974, 1976) y LOGIST (Wood, Wingersky y Lord, 1976).

Por otro lado, se produce una explosión de trabajos que difunden las aplicaciones prácticas, y la comunidad científica y profesional comienza a ver su utilidad. En las principales revistas se dedican números monográficos a presentar la vertiente aplicada de la teoría, por ejemplo, el *Journal of Educational Measurement* (1977, Vol. 14, Nº. 2), *Applied Psychological Measurement* (1982, Vol. 6, Nº. 4) o el *International Journal of Educational Research* (1989, Vol. 13, Nº. 2), entre otras. Asimismo se publican los manuales y monogramas de Andrich (1988), Baker (1985), Hambleton y Swaininathan (1985), Hambleton, Swaminathan y Rogers (1991), Hulin, Drasgow y Parsons (1983), Lord (1980), Weiss (1983) y, más recientemente, el de Fisher y Molenaar (1995), sobre los últimos desarrollos con el modelo de Rasch, y el de van der Linden y Hambleton (1997) sobre los últimos avances en general, y el de Embretson y Hershberger (1999) sobre las nuevas reglas de medición. También hay que señalar las revisiones realizadas por Goldstein y Wood (1989), Hambleton (1990b, 1994b), Hambleton y Rogers (1991), McKinley (1989), Muñiz y Hambleton (1992) o Traub y Lam (1985). En nuestro país, el proceso es más lento, pero también le va llegando su hora con los libros de López-Pina (1995), Martínez-Arias (1995), Muñiz (1990, 1996b, 1997b), Santisteban (1990) y Tomás, Oliver y Meliá (1992). En la actualidad sigue esta misma tónica con un volumen creciente de publicaciones en los diversos temas especializados que constituyen la TRI.

En general, describir las características generales de la TRI, dada la diversidad y complejidad de las aportaciones en los últimos años, resulta difícil de realizar; no obstante, vamos a presentar, en el apartado siguiente, los supuestos fundamentales de esta teoría que nos permitirá una aproximación más comprensiva a los modelos y variantes de la misma.

1.2. Supuestos fundamentales.

El objetivo principal de la TRI es *"conseguir medidas invariantes respecto de los sujetos medidos y de los instrumentos utilizados"* (Muñiz, 1997, p.17). En la unificación de estos dos conceptos, separación de parámetros e invarianza de los mismos, está la clave del éxito de esta teoría. Para conseguir estos objetivos, la TRI desarrolla un

conjunto de modelos matemáticos que comparten esta idea central, es decir, todos asumen que la probabilidad de que una persona emita una determinada respuesta ante un ítem puede ser descrita en función de la posición de la persona en el rasgo o aptitud latente (variable que suele denominarse genéricamente con la letra griega θ) y de una o más características de ítem (índice de dificultad, de discriminación, probabilidad de acertar por azar,..) . Es por ello, que los principales supuestos de esta teoría son proposiciones referidas a la naturaleza del rasgo que se pretende medir *(supuesto de unidimensionalidad del espacio latente*) y a las relaciones que se esperan entre las respuestas de los items (*independencia local*).

1.2.1. Unidimensionalidad del espacio latente.

La TRI propone modelos matemáticos que describen las respuestas de los sujetos en función de su localización en el rasgo latente. En consecuencia, una de las primeras cuestiones que debe resolver se refiere a la naturaleza de ese rasgo, y en concreto, al número de componentes o variables latentes que es necesario tomar en consideración para describirlo adecuadamente. Así, la dimensionalidad del espacio latente quedaría definida cuando se identifican esos componentes. Cuando esto sucede, *"la distribución condicional de la puntuación del ítem para un θ fijado es la misma para todas las poblaciones de interés"* (Lord y Novick, 1968, p. 359). Tal como ocurre en el modelo de regresión lineal, esa distribución condicional es independiente de la distribución del rasgo en la población. Este hecho es de gran importancia, ya que introduce en la TRI la posibilidad de controlar la variación sistemática de las puntuaciones de los sujetos debida a su posición en el rasgo. Efectivamente, esa variabilidad puede eliminarse de la distribución de probabilidad de acierto de un ítem simplemente condicionando esa distribución a un valor fijo del rasgo. En consecuencia, las distribuciones de probabilidad de acierto de items diferentes condicionadas a un mismo valor en el rasgo[1] mostraran una única fuente de variabilidad sistemática: la que introducen los items. Más adelante veremos cómo este hecho constituye uno de los pilares sobre los que se asienta la propiedad de invarianza de parámetros que caracteriza a esta teoría.

No obstante, el cumplimiento del supuesto de unidimensionalidad del espacio latente, en sentido estricto, es difícil de mantener. Asumir que en la práctica es posible definir todos los componentes de los que depende del comportamiento manifiesto de los sujetos, aunque esa conducta sea aparentemente tan sencilla como acertar o fallar un ítem, es difícilmente sostenible. Hambleton y Swaminathan, (1985) señalan que en la práctica, la definición del espacio latente se limita a la exigencia de una *"dimensión dominante"*, es decir, basta que exista un rasgo principal que sea dominante o relevante para discriminar entre grupos de examinados. En cualquier caso, la aplicación de un modelo de TRI a un conjunto de datos no depende de la opinión acerca de la dimensionalidad del rasgo de que se trate, sino que se fundamenta en la comprobación empírica del grado en que los datos satisfacen este supuesto. La lógica que subyace a esta comprobación empírica puede encontrase en Santisteban (1995) y Muñiz, (1997).

Hidalgo (1998) presenta una revisión de los diferentes métodos empleados para detectar la unidimensionalidad, recogiendo desde los métodos más clásicos basados en diferentes variantes del análisis factorial, hasta los más recientes, como el procedimiento DETECT (Zhang y Stout, 1997). En el trabajo de Cuesta (1996) también podemos encontrar una revisión detallada de estos métodos y además recoge la problemática asociada al uso

[1] En la literatura se suele reducir este concepto con la expresión "distribuciones de probabilidad condicionadas". En adelante se utilizará también en este texto siguiendo el modo usual, es decir, sobreentendiendo que la expresión se refiere a la probabilidad de acierto, y que están condicionadas a un mismo valor en el rasgo.

del análisis factorial en el contexto de la TRI: los modelos de la TRI suelen aplicarse 1) a datos dicotómicos, y 2) asumiendo que la relación respuesta-rasgo latente no es lineal. Estos dos aspectos entran en conflicto directo con la técnica del análisis factorial lineal (Carroll, 1988 y Maydeu ,1996). Estos autores también exponen las diferentes soluciones adoptadas, entre las que figuran el análisis factorial no lineal propuesto por McDonald e implementado en el programa NOHARM (Fraser y McDonald, 1988), el método propuesto por Muthén e implementado en el programa LISCOMP (Muthén, 1987), y el análisis factorial de información completa propuesto por Muraki e implementado en los programas TESTFACT (Wilson, Wood, y Gibbons, 1993) y POLYFACT (Muraki, 1993).

En resumen, la mayor parte de los modelos desarrollados bajo esta teoría asumen la unidimensionalidad del espacio latente. La razón es clara, siendo mucho más sencillo explicar la probabilidad de dar una determinada respuesta a un ítem en función de una variable que no en función de varias, por cuanto que estimar la influencia de cada una de ellas sobre la misma respuesta es un asunto complicado.

1.2.2. Independencia local.

La información en que se basa la TRI para estimar los parámetros de los sujetos procede únicamente de los patrones de respuestas de los mismos. De ahí que sea necesario imponer ciertas restricciones sobre las relaciones que puedan aparecer entre las respuestas en un patrón dado, y también entre patrones de respuestas. La independencia local de las respuestas es la restricción más importante. Ésta consiste en asumir que las respuestas de diferentes sujetos j con un determinado nivel i en el rasgo (θ_{i1}, θ_{i2},...,θ_{ij}) a un ítem son también estadísticamente independientes de las respuestas de esos sujetos a cualquier otro ítem, es decir, cada nueva respuesta es independiente de la respuesta anterior, y éstas sólo vienen determinadas por la probabilidad de acierto a ese ítem, que para sujetos con igual aptitud, es la misma para todo el grupo.

En el contexto definido por las respuestas de un conjunto de sujetos con igual puntuación en el rasgo a un conjunto de items, la independencia entre las respuestas de los sujetos sólo aparece localmente, es decir, una vez eliminado el efecto del nivel de aptitud de los mismo. En estos casos la frecuencia relativa de aciertos de los sujetos con ese nivel en el rasgo en el ítem k (f_{ik}) es la misma, tanto si se consideran todos los sujetos con ese nivel en el rasgo, como si se separan en función del acierto o fallo a otro ítem cualquiera t (f_{ik} [x_{ik} |X_{it} =1] = f_{ik} [x_{ik} |X_{it} =0]= f_{ik}). Cuando se considera el conjunto de respuestas de sujetos con diferentes niveles de aptitud, las respuestas de estos sujetos presentaran la relación habitual, es decir, una relación positiva entre el nivel en el rasgo y la probabilidad de acierto.

Comparando este supuesto con el de la unidimensionalidad del espacio latente, podemos apreciar que ambos supuestos garantizan la independencia de las distribuciones condicionales, sólo que en un caso respecto del grado de aptitud de los sujetos, y en el otro respecto de las respuestas dadas a otros items. El supuesto de la unidimensionalidad del espacio latente garantiza la independencia de la distribución de probabilidad condicional de un ítem a través de los diferentes niveles de aptitud, y, por tanto, elimina la dependencia de la distribución de la aptitud en la población. El supuesto de independencia local garantiza la independencia de la distribución de probabilidad condicional de un ítem respecto de las respuestas dadas a otros items, y, por tanto, elimina la dependencia del test.

Como consecuencia de ello, la comparación de las distribuciones de probabilidad condicionada de diferentes items permite atribuir las diferencias detectadas, entre esas distribuciones, únicamente a las diferentes características de los items, posibilitando así su estimación. Este es el modo en que la TRI responde a las consideraciones de Thurstone referentes a que el objeto al que se aplica la medición no debe influir mucho sobre la función propia del instrumento de medición, y las de Lord cuando expresaba en los siguientes términos el mismo deseo "debemos poder estimar la aptitud de un examinado a partir de cualquier conjunto de items que le puedan ser administrados" (Lord, 1980, p.11)

En resumen, la consecuencia inmediata de este supuesto así definido es que la probabilidad de un sujeto con aptitud θ, obtenga un determinado patrón de respuestas en un conjunto de items ($i=1, 2, 3,...,n$) es igual al producto de las probabilidades de respuesta a cada uno de los items condicionadas a ese nivel de aptitud; formalmente puede expresarse como sigue:

$$P(X_1, X_2, X_3,..., X_n \mid \theta) = \prod_{i=1}^{n} P(X_i \mid \theta) \qquad (1.1)$$

donde X_i es la respuesta de un sujeto al ítem i.

Este hecho permitirá, una vez conocidas las características de los items, estimar el nivel de aptitud de cada sujeto en función del patrón de respuestas particular que presente, sin necesidad de comparar ese patrón con los presentados por el resto de sujetos de la población. Con ello se resuelve otro importante problema planteado en la Teoría Clásica de Tests, ya que las puntuaciones de los sujetos no han de ser referidas a ningún grupo normativo para poder proceder a su interpretación.

La lógica de la contrastación empírica de este supuesto puede encontrarse en Santisteban (1990), aunque no es necesario proceder a su contrastación dado que la adecuada especificación de la unidimensionalidad del espacio latente lleva aparejada la independencia local de los items, es decir, los supuestos de unidimensionalidad e independencia local están relacionados de forma asimétrica, esto es, cuando hay unidimensionalidad también existe independencia local, pero no al contrario. Por este motivo, no se ha prestado mucha atención a la comprobación del supuesto de independencia local, ya que basta con demostrar el de unidimensionalidad, y el esfuerzo se ha concentrado sobre éste.

Los supuestos de unidimensionalidad e independencia local son los supuestos centrales y característicos de la TRI. Pero también se asumen otros supuestos de carácter más general.

1.2.3. Otros supuestos.

El primero de estos supuestos hace referencia a la naturaleza continua del espacio latente definido por el rasgo. Los modelos en que se asume que el espacio latente definido por el rasgo es discreto se estudian bajo la Teoría de la Clase Latente, que junto con la TRI se enmarcan dentro de la Teoría del Rasgo Latente (Lazarfeld, 1950). Desarrollos más actuales sobre los modelos de clases latentes y algunas de sus aplicaciones prácticas más recientes pueden encontrase en Clogg (1981); McCutcheon, (1987); Douglas, (1988); Hagenaars, (1990, 1993) y Rost y Langeheine (1997).

El segundo de estos supuestos asume que la probabilidad de dar la respuesta correcta a un ítem aumenta a medida que se incremento el nivel de aptitud. A este supuesto se le denomina supuesto de monotonicidad simple. Este supuesto se cumple si y sólo si, dados dos sujetos cualesquiera, ***j*** y ***k***, tal que $\theta_j > \theta_k$,. Entonces $P(\theta_j) > P(\theta_k)$, cualquiera que sea el ítem empleado para verificar esa relación. Por lo tanto, este supuesto garantiza que cada ítem considerado de forma aislada ordena del mismo modo a todos los sujetos. También existe una modalidad de monotonicidad más restrictiva, denominada doble monotonicidad, que exige que, dados dos items ***j*** y ***k*** y un mismo nivel de aptitud (θ_i), si $P_j(\theta_i) > P_k(\theta_i)$, entonces $P_j(\theta) > P_k(\theta)$, para cualquiera que sea el valor de la aptitud considerado. Este supuesto garantiza que la ordenación de los items, establecida en función de la probabilidad de acierto de los sujetos, es la misma cualquiera que sea el valor del rasgo considerado.

Por último, una asunción que Hambleton y Swaminathan (1985) consideran que está implícita entre los supuestos de la TRI se refiere a que los tests utilizados para ajustar los diferentes modelos no sean administrados bajo condiciones de velocidad, es decir, la teoría plantea que cuando un sujeto falla un determinado items se debe a que su nivel de habilidad "está limitado" para responder adecuadamente al mismo y no por falta de tiempo para llegar al ítem. Esta cuestión está ligada al supuesto de unidimensionalidad. Dado que cuando un test de ejecución está afectado por la velocidad debemos considerar que existen dos rasgos implicados en el mismo: la habilidad que mide el test y la velocidad de ejecución de la tarea.

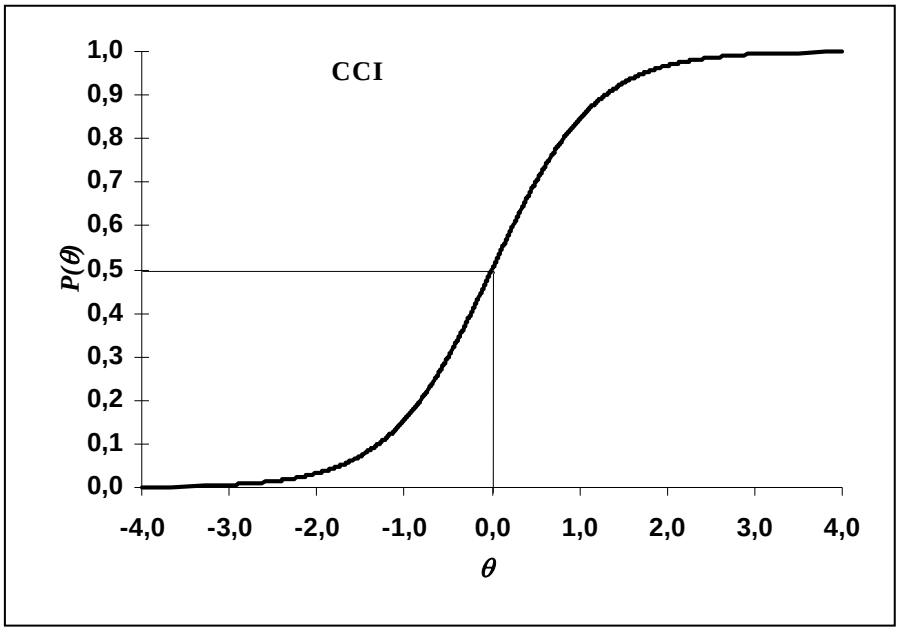
En resumen, el conjunto de estos supuestos permite llevar a la práctica la idea central de la TRI, es decir, expresar la probabilidad de que una persona emita una determinada respuesta ante un ítem, generalmente la respuesta correcta o positiva, en función de la posición de la persona en el rasgo latente y de una o más características del ítem. Como expresan Hulin, Drasgow y Parsons (1983), los modelos de TRI proporcionan una estrategia probabilística para trabar o enlazar las respuestas de los sujetos -las variables observables- con los constructos teóricos contenidos en las teorías psicológicas -los rasgos latentes-. Las curvas características de los items constituyen el elemento de enlace como veremos a continuación.

1.2.4. Curva Característica del ítem.

Como ya hemos señalado, existe una relación funcional entre los valores de la variable que miden los items y la probabilidad de acertar éstos, y de ahí que un objetivo de la TRI sea establecer la mejor función que ajuste esta relación, es decir, una función que de cuenta de la relación entre la probabilidad de acertar el ítem con la localización en el rasgo de los sujetos; en concreto, esa relación puede ser expresada mediante una función (ver figura 1.1) de regresión no lineal que une cada valor en el rasgo con la puntuación medida condicionada en el ítem, que, en el caso de items dicotómicos, coincide con la probabilidad condicionada al nivel de θ de acertar el ítem.

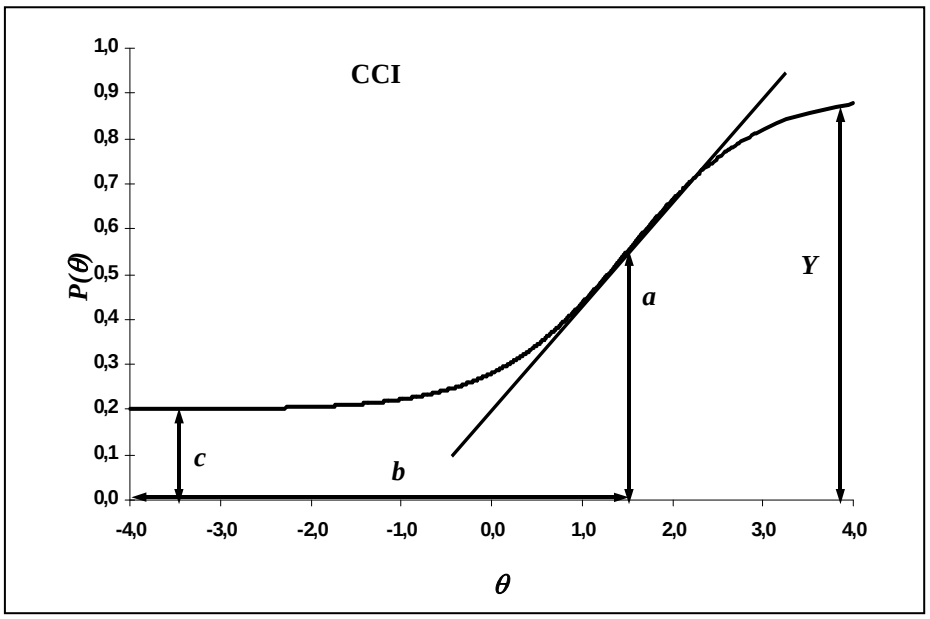
Esta función de enlace recibe el nombre de *Curva Característica del ítem* **(CCI)**, *Huella del ítem* o *Función de Respuesta al Ítem* (Lord y Stocking, 1988, Fischer, 1995). Cada ítem está caracterizado por una CCI particular y propia, es decir, las CCI de los items que miden una determinada variable θ no son iguales, aunque comparten determinadas formas generales como se verá mas adelante. La forma particular de cada CCI depende de los parámetros o características de cada ítem. Como se ha apuntado, bajo esta teoría las características del ítem son independientes de la distribución de la aptitud en la población de sujetos. Lo que implica que la CCI también es igualmente invariante o independiente del objeto medido. Y, como Lord y Novick (1968) indican, dado que relaciona

la localización de los sujetos en el rasgo con sus respuestas, también es útil para realizar inferencias en el sentido opuesto, es decir, de las respuestas de los sujetos a su localización en el rasgo, que es el objetivo del proceso de medida.



Figura, 1.1.- Curva característica de un ítem.

Aunque las características de los items pueden ser numerosas, en particular son tres los parámetros que se suelen proponer[2] para la obtención de las CCIs, (véase la figura 1.2):



Figura, 1.2.- Parámetros de la curva característica de un ítem.
(*θ N(0,1)* y, ***a***=1, ***b***=1.5, ***c***=0.2 e ***Y***=.9*)*

[2] Algunos autores (Bartón y Lord, 1981) han propuesto un cuarto parámetro, ***Y***, para caracterizar aquellos items en los que sujetos con alta competencia fallan el ítem impropiamente. Suele tomar valores ligeramente inferiores a 1(Muñiz, 1997).

Parámetro a

El parámetro ***a*** se le denomina ***índice de discriminación*** del ítem, y representa la magnitud del cambio en la probabilidad de acertar el ítem conforme varía el nivel de habilidad. Su valor es proporcional a la pendiente de la recta tangente a la CCI en el punto de inflexión de ésta.

Parámetro b

El parámetro ***b*** se corresponde con el valor en la abscisa (escala de habilidad (θ)) del punto de máxima pendiente de la CCI. Se le denomina ***índice de dificultad*** del ítem, y es un parámetro de localización del ítem que representa la posición de la CCI en relación al nivel de habilidad (θ) necesario para obtener una probabilidad de acierto $P(\theta)=(1+c)/2$.

Parámetro c

El parámetro ***c*** es el índice de pseudo-azar del ítem, representa la probabilidad de acertar de los sujetos que desconocen la respuesta correcta, es decir, es el valor de $P(\theta)$ cuando θ tiende a su valor mínimo ($-\infty$).

La CCI queda definida cuando se especifican estos tres parámetros y se adopta una determinada función matemática para conformar la curva. Según el tipo de función matemática adoptada, el número de parámetros referidos a los ítems considerados como relevantes, el tipo de respuesta y la dimensionalidad del espacio latente obtendremos diferentes modelos o tipos de CCI. En el siguiente apartado nos referiremos a las funciones de enlace o modelos mas utilizados en esta teoría.

1.3. Modelos Básicos.

Una de las primeras funciones de enlace propuestas en el contexto de la TRI fue la función de regresión lineal (Lazarsfeld, 1959). Ésta fue abandonada pronto porque presentaba limitaciones respecto a los planteamiento de la TRI, en concreto, para determinados valores de θ, la probabilidad $P(\theta)$ podría ser negativa o mayor que 1. Lo que se necesitaba era encontrar una función matemática tal que fuera monótona creciente en θ, como la de regresión lineal, pero que presentara valores asintóticos superiores en torno a 1 e inferiores en torno a 0. Esta limitación orientó la búsqueda hacia diferentes transformaciones de la probabilidad de respuesta que cumplieran estas condiciones. Como señala Muñiz (1997), hasta la actualidad se utilizan dos tipos de familias de funciones de distribución para la CCI: la función logística y la curva normal acumulada, que dan lugar a seis modelos generales según se contemple uno, dos o tres parámetros de los items para cada una de estas dos funciones. En todos los casos son modelos unidimensionales y asumen que la respuesta a los items es dicotómica[3]. El desarrollo histórico de estos modelos podemos enmarcarlo en los trabajos de Lord y su modelo de ojiva normal, Rasch y su modelo logístico de un parámetro y Birnbaum y sus modelos logísticos de 2 y 3 parámetros.

1.3.1. La contribución de Lord: el modelo de ojiva Normal.

Los primeros modelos asumidos por la TRI proponen la función de la curva normal acumulada para la expresión de la CCI. Se denominaron modelos de ojiva normal y precedieron en su desarrollo a los modelos

[3] Para modelos multidimensionales y modelos de respuestas politómicas o multicategoriales existen en la literatura otros tipos de modelos que abordaremos en el apartado dedicado a los desarrollos de los modelos básicos.

logísticos. Fueron sugeridos por Thurstone (1928) y desarrollados por Richardson (1936), Lawley (1943), y Tucker (1946), hasta recibir su formulación definitiva por parte de Lord (1952).

Lord desarrolló el modelo de ojiva normal de dos parámetros. Estos dos parámetros eran la dificultad del ítem y su discriminación, y fueron incluidos en el modelo debido a que eran las dos características de los items tradicionalmente estudiadas desde la Teoría Clásica de Tests. Como se ha señalado, estos dos parámetros están estrechamente ligados al origen de la TRI, ya que fue su dependencia de la población bajo la Teoría Clásica de Tests lo que motivó la búsqueda de teorías alternativas. Este modelo puede ser expresado de la siguiente forma:

$$\boldsymbol{P_i(\theta) = P(X_i = 1 \mid \theta) = \Phi(a_i(\theta - b_i))} \qquad (1.2)$$

donde $\boldsymbol{\Phi}$ representa a la función de ojiva normal, y los parámetros $\boldsymbol{a_i}$, $\boldsymbol{b_i}$, y $\boldsymbol{\theta}$ mantienen el significado ya conocido.

La siguiente, es una formulación general del modelo de ojiva normal, atendiendo a los cuatro parámetros que pueden utilizarse (tomado de Muñiz, 1997):

$$P_i(\theta) = c_i + (Y_i - c_i)\int_{-\infty}^{a_i(\theta - b_i)} \left(\frac{1}{\sqrt{2\pi}}\right) . e^{-\left(\frac{z^2}{2}\right)\partial_z} \qquad (1.3)$$

donde los parámetros $\boldsymbol{a_i}$, $\boldsymbol{b_i}$, $\boldsymbol{c_i}$, $\boldsymbol{Y_i}$ y $\boldsymbol{\theta}$ mantienen el significado anteriormente expuesto.

Si el parámetro $\boldsymbol{Y_i}$ toma un valor igual a **1,** la expresión anterior se convierte en el modelo para tres parámetros [$\boldsymbol{P_i(\theta)=c_i}$ +(1- $\boldsymbol{c_i}$) Φ($\boldsymbol{a_i}$ ($\boldsymbol{\theta}$ - $\boldsymbol{b_i}$))] y, en el modelo de dos parámetros [$\boldsymbol{P_i(\theta)}$=Φ($\boldsymbol{a_i}$ ($\boldsymbol{\theta}$ - $\boldsymbol{b_i}$))] cuando $\boldsymbol{c_i}$ toma un valor igual a **0**. Del mismo modo, si el parámetro $\boldsymbol{a_i}$ toma un valor igual a **1,** la expresión se convierte en el modelo de un parámetro [$\boldsymbol{P_i(\theta)}$= Φ($\boldsymbol{\theta}$-$\boldsymbol{b_i}$)].

En resumen, el modelo de ojiva normal fue el primer modelo susceptible de ofrecer la propiedad de invarianza de parámetros, aplicable tanto a los parámetros de los items como a los de los sujetos. Y como tal marcó todo un hito en la historia de la Psicometría. Sin embargo, presentaba dos aspectos dificultosos: su complejidad matemática y la estimación práctica de sus parámetros. De hecho este último problema detuvo la aplicación del modelo durante varios años, y no fue resuelto hasta 1968, año de la publicación del libro de Lord y Novick. En él, los autores ofrecían un método de estimación de los parámetros de los items basado en elementos de la Teoría Clásica de Tests como la proporción de aciertos y la correlación ítem-test. Pero a pesar de ello, lo cierto es que el modelo continuaba siendo matemáticamente difícil de tratar, y las estimaciones de sus parámetros seguían presentando problemas, por lo que apenas se llegaron a formular pruebas de bondad de ajuste asociadas a este modelo.

1.3.2. La contribución de Rasch: el modelo logístico de Rasch.

El concepto de medida objetiva, tal como lo definió Thurstone, fue durante años el objetivo de la Teoría Psicométrica. Es por ello que el propósito inicial de Rasch, probablemente, pudo ser la búsqueda de la invarianza, en el sentido empleado por Thurstone casi tres décadas antes, y terminó siendo la elaboración de un modelo de medida. Su *"modelo estructural para los items de un test"* -que es como lo denominaba- fue formulado desde una perspectiva original. Rasch no apeló a los desarrollos de Lawley o Lord en la fundamentación de su modelo (Hambleton, 1994;Andersen, 1995). En este contexto, su primer logro fue la introducción del modelo poissoniano donde mostró

que, bajo ciertos supuestos, es posible asignar valores escalares a las dificultades relativas de un conjunto de tests, siendo esas estimaciones de la dificultad independientes de la población de sujetos estudiada (el de Rasch). Las claves de este primer logro estaban en: a) la aplicación de un modelo probabilístico (el modelo de Poisson), b) la expresión del parámetro correspondiente a este modelo (probabilidad de cometer un error, $P(e)$, al contestar un ítem en función de otros dos parámetros: la destreza del sujeto, ξ_i, y la dificultad del test, δ_j, donde $P(e_{ij}) = f(\delta_i / \xi_j)$), c) la disponibilidad de estimadores suficientes de esos dos parámetros, y, d) en el uso del concepto de probabilidad condicional aplicado a un diseño multimuestra-multitests.

Rasch reformuló el parámetro del modelo de Poisson en los términos indicados anteriormente, y resultó que, bajo ciertas condiciones, el modelo resultante proporciona estimadores suficientes de los parámetros, lo que permitió por primera vez la estimación separada e independiente de la dificultad del test y de la aptitud de los sujetos. Pero este modelo presentaba serias limitaciones: 1) no podía aplicarse a tests aislados, y 2) solo se podía ajustar a tests formados por items de igual dificultad. Esta última restricción era particularmente difícil de satisfacer por la mayor parte de los tests de aptitudes. En consecuencia, Rasch pensó en un nuevo modelo que incluyera un parámetro representativo de la dificultad de cada uno de los items, en lugar de un solo parámetro para la dificultad del test. Este modelo debía contener un parámetro para cada sujeto (ξ) y un parámetro para cada ítem (δ). A continuación había que seleccionar un indicador que fuera función de esos dos parámetros. Rasch eligió como indicador la probabilidad de dar la respuesta correcta a un ítem. Y aplicando un razonamiento análogo al empleado en el modelo anterior, Rasch propuso que la probabilidad de que un sujeto acertara un ítem $P(X_i=1)$ debía de ser función del cociente entre los dos parámetros:

$$P(X_i = 1) = \xi / \delta \tag{1.4}$$

Finalmente quedaba proponer la función de enlace. Y la función más simple de entre las que crecen de 0 a 1 a medida que ξ/δ tiende a infinito es:

$$f(X) = \frac{(\xi / \delta)}{1 + (\xi / \delta)} \tag{1.5}$$

más conocida en su versión logística, donde $\log \xi = \theta$ (la puntuación en el rasgo), y $\log \delta = b$ (la dificultad del ítem):

$$P(X_{ij} = 1 \mid \theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \tag{1.6}$$

que suele expresarse de forma simplificada como:

$$P(X_{ij} = 1 \mid \theta) = \frac{1}{1 + e^{-(\theta - b_i)}} \tag{1.7}$$

o alternativamente como:

$$P(X_{ij} = 1 \mid \theta) = (1 + \exp - (\theta - b_i))^{-1} \tag{1.8}$$

El modelo logístico tiene la ventaja de que, mediante el uso de una constante adicional, (D=1.7) sus valores se aproximan notablemente a la curva normal acumulada, por lo que es frecuente encontrarla expresada como función logística normalizada. En este caso suele recibir el nombre de función log-normal . Su expresión es la siguiente:

$$P(X_{ij} = 1 \mid \theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}} \tag{1.9}$$

y alternativamente,

$$P(X_{ij} = 1 \mid \theta) = \frac{1}{1 + e^{-D(\theta - b_i)}} \tag{1.10}$$

ó

$$P(X_{ij} = 1 \mid \theta) = (1 + \exp - D(\theta - b_i))^{-1} \tag{1.11}$$

Una vez formulado el modelo, el siguiente paso que había que resolver era el relacionado con las estimaciones de los parámetros. Rasch (1960) propuso tres métodos de estimación de parámetros, el método **LOG**, el método **PAIR**, y el método **FCON**. Estos métodos quedaron obsoletos tras la aplicación de la teoría de máxima verosimilitud a la estimación de los parámetros de este modelo. En la actualidad son estos métodos máximo-verosímiles los más utilizados.

En resumen, el modelo de Rasch (1960) es formalmente parte de la familia de los modelos logísticos desarrollados por Birnbaum (1968) que expondremos en el apartado siguiente.

1.3.3. La contribución de Birnbaum: los modelos logísticos.

El propósito de Birnbaum (1968) fue resolver los inconvenientes que presentaba el modelo de ojiva normal de Lord. Se hacía necesario encontrar una alternativa a ese modelo que conservara todas sus ventajas, pero que resultara matemáticamente más tratable. En efecto, su propuesta fue sustituir la función de ojiva normal por la función logística, dando lugar a la familia de modelos que se aplican en la actualidad.

Birnbaum (1957,1958,1968) desarrolló un modelo logístico de dos parámetros puede ser expresado como sigue:

$$P_i(\theta) = P(X_i = 1 \mid \theta) = \eta(a_i(\theta - b_i)) \tag{1.12}$$

donde η representa a la función logística, y los parámetros a_i, b_i, y θ mantienen también el significado anteriormente expuesto.

La consideración de un tercer parámetro se debe a Birnbaum, que propuso la incorporación del parámetro c que, como ya hemos señalado, proporciona la asíntota mas baja para la curva $Pi(\theta)$ y que suele interpretarse como la probabilidad de que examinados con niveles bajos de aptitud respondan correctamente, es decir, la probabilidad de acertar por azar.

La siguiente es una formulación general del modelo logistico normal, atendiendo a los cuatro parámetros que suelen utilizarse (tomado de Muñiz, 1997):

$$P_i(\theta) = c_i + (Y_i - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \tag{1.13}$$

y sus fórmulas alternativas,

$$P_i(\theta) = c_i + (Y_i - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}} \tag{1.14}$$

ó

$$P(\theta) = c_i + (Y_i - c_i)(1 + \exp - D(\theta - b_i))^{-1} \tag{1.15}$$

donde los parámetros a_i, b_i, c_i, Y_i y θ mantienen el significado anteriormente expuesto.

Si el parámetro Y_i toma un valor igual a **1,** la expresión anterior se convierte en el modelo para tres parámetros [$P_i(\theta)=c_i +(1- c_i)\ \eta(Da_i\ (\theta - b_i\ ))$] y en el modelo de dos parámetros [$P_i(\theta)=\eta\ (Da_i\ (\theta - b_i\ ))$], cuando c_i toma un valor igual a **0**. Del mismo modo, si el parámetro a_i toma un valor igual a **1,** la expresión se convierte en el modelo de un parámetro [$P_i(\theta)=\eta(D(\theta -b_i)$)].

Pero la contribución de Birnbaum no se limitó a la formulación de modelos, sino que también desarrolló una técnica de estimación de parámetros que, aunque sustancialmente mejorada, todavía sigue vigente. Se trata del método de estimación de máxima verosimilitud conjunta (**JML**).

A grandes rasgos, este método procede a obtener los estimadores de máxima verosimilitud de los parámetros de los items y de los sujetos de forma conjunta e iterativa, comenzando por fijar los valores de los parámetros de los sujetos a aquellos valores iniciales que se consideren más adecuados, para proceder a la obtención del primer conjunto de estimaciones de los parámetros de los items. A continuación mantiene los parámetros de los items fijos a esos valores y estima los valores de los parámetros de los sujetos. Entonces se comienza de nuevo el proceso, pero empleando los nuevos estimadores de los parámetros de los sujetos, y así sucesivamente. Esta secuencia continúa en un proceso iterativo hasta que se alcanza la convergencia, de acuerdo con el criterio que se haya establecido de antemano. El mérito de este método estriba en el hecho de que es capaz de obtener estimaciones de los parámetros implicados en el modelo cuando no es posible separar las ecuaciones destinadas a la estimación de los parámetros de los items y de los sujetos.

Finalmente, otro de los aspectos propuesto por Birnbaum es el relacionado con la cantidad de información proporcionada por los items y el test respecto del nivel de habilidad de los sujetos. Birnbaum introdujo la medida de información de Fisher en el contexto de la información proporcionada por un test. Esta información viene dada por la función de información del test, y presenta la siguiente expresión:

$$I(\theta) = \Sigma I_i(\theta) \qquad (1.16)$$

donde $I_i(\theta)$ es la información del ítem i condicionada a θ. Esta expresión indica la precisión de las puntuaciones que ofrece el test condicionada a cada uno de los valores que puede tomar la aptitud bajo estudio. Esta nueva aproximación al concepto de fiabilidad de las puntuaciones vino a resolver otro de los grandes inconvenientes de la Teoría Clásica de los Tests: el de la homocedastidad del error de medida a lo largo de toda la distribución de valores de la aptitud. También fue Birnbaum quien sugirió la utilidad de esta función de información en la construcción de tests.

Resumiendo, podemos decir que Lord propuso un modelo en el que por primera vez la medida objetiva se convertía en una realidad. La aportación de Rasch no cifra su valor en la utilidad del modelo que propuso, más importante que su utilidad es el puente que su modelo establece entre la Teoría de los Tests y la Teoría de la Medición. Dos caras de una moneda que hasta ese momento parecían irreconciliables. Birnbaum simplificó ese modelo y proporcionó un método de estimación de parámetros, lo que revirtió en una mayor aplicabilidad de ese modelo. Y, además, vio la utilidad del procedimiento de Fisher para evaluar la precisión del test, e incorporó esta información al proceso de construcción de un test.

Todo ello puso de manifiesto el potencial de estos modelos a la hora de resolver algunos de los problemas que la Teoría Clásica de Tests había dejado planteados, y contribuyó a mostrar que el incremento en complejidad que presentaba la TRI respecto de la Teoría Clásica de Tests era, cuanto menos, proporcional al incremento en su utilidad. Este hecho fue decisivo en la expansión de la nueva teoría.

Estos modelos constituyen la primera generación de modelos desarrollados bajo la TRI. Desde 1968 hasta la actualidad el progreso que ha experimentado esta teoría ha sido incesante y en distintas direcciones. Una de estas direcciones corresponde al desarrollo de nuevos modelos, más flexibles en unos casos, más específicos en otros, pero siempre tratando de responder a la gran variedad de situaciones en que es necesario medir en Psicología. A continuación vamos a presentar los principales desarrollos y algunas de las clasificaciones mas significativas.

1.4. Desarrollo de los modelos básicos y su clasificación.

Las extensiones que se han desarrollado a partir de los primeros modelos logísticos de 1, 2, y 3 parámetros son muy numerosas y variadas. Vamos a tratar de describir brevemente estas extensiones resaltando qué nuevas posibilidades de medida proporcionan, cuáles son las líneas generales en que se basa la estrategia metodológica que permite esos avances, y cuáles son los modelos más representativos en cada caso.

Extensiones del modelo de Rasch.

Una de las líneas de desarrollo del modelo de Rasch más importantes es la liderada por Fischer. El objetivo de esta línea de desarrollo consiste en flexibilizar el modelo de Rasch de forma que permita descomponer la aptitud que se supone responsable de la respuesta del sujeto en sus componentes cognitivos más básicos. El modelo que desarrolla esta idea es el Modelo de Tests Lineal Logístico (Linear Logistic Test Model (**LLTM**), Fischer, 1973, 1995, 1997, 1998), y la estrategia central en que se basa, consiste en el modelado del parámetro de dificultad del ítem como una función de la dificultad de las operaciones cognitivas implicadas en la resolución del ítem. Este modelo asume que el parámetro de dificultad puede ser descompuesto en la suma ponderada de determinados parámetros básicos, que representan la dificultad de cada una de esas operaciones cognitivas. Este modelo queda pues limitado a aquellos tests en que se cuente con suficiente información acerca de cuáles son los componentes cognitivos implicados en la respuesta correcta a cada ítem.

Fischer también presenta otra variante de este modelo, el Modelo de Tests Lineal Logístico para el Cambio (Linear Logistic Test Model for Change (**LLTMC**), Fischer, 1995, 1997, 1998). Este modelo constituye una reinterpretación del modelo anterior aplicable al caso en que cada ítem es presentado en dos o más ocasiones a cada sujeto, con el objetivo de evaluar el cambio entre las aplicaciones del test. En este caso las diferentes aplicaciones del test dan lugar al desdoblamiento de cada ítem en diferentes items "virtuales" (porque en realidad son el mismo). La idea central se origina en el hecho de que cualquier cambio en el parámetro de los sujeto al contestar a un ítem en diferentes ocasiones puede ser descrito sin perdida de generalidad (y bajo el modelo de Rasch) como un cambio en la dificultad de ese ítem. El ítem se vuelve más "fácil" o "difícil" para los sujetos. Ese cambio es un indicador del cambio experimentado por los sujetos en el atributo evaluado. La dificultad del ítem virtual puede descomponerse en la combinación aditiva del componente "dificultad inicial" más el componente "cambio inducido entre las aplicaciones del ítem". Esta estrategia permite aplicar el modelo de Rasch manteniendo todas sus ventajas a aquellas situaciones en que se pretenda evaluar el cambio, ya sea por efecto de un tratamiento, de un entrenamiento, etc.

También existe una versión multidimensional de este modelo denominada modelo lineal logístico de supuestos relajados (Linear Logistic Test Model for Relaxed Assumptions (**LLRA**), Fischer, 1995; Fischer y Seliger, 1997).

Los Modelos de Interacción y de Aprendizaje (Interaction Models and Learning Models, Verhlst y Glas, 1995) representan otra línea de desarrollo relacionada con la anterior, en el sentido de que son modelos que también se aplican en los casos en que se produce algún cambio en el parámetro de los sujetos. La diferencia estriba en que en los modelos anteriores ese cambio se produce entre las aplicaciones del test, mientras que en estos modelos el cambio se produce durante la aplicación del test. En otras palabras, el sujeto "aprende" o cambia su opinión (p.e. modifica su probabilidad de "acierto" a los items) a medida que va realizando el test. En este caso existen modelos que no asumen el supuesto de independencia local, como los modelos log-lineales de Kelderman (1984, 1997) y los modelos conjuntivos de Jannarone (1986, 1997). Estos modelos, que ya no son modelos de TRI porque no modelan la respuesta al ítem sino al test, pueden ser considerados como casos particulares del Modelo Interactivo desarrollado por Verhlst y Glas (1995). También existen modelos que sí conservan este supuesto, como el Modelo de Aprendizaje de Verhlst y Glas (1995), en el que se emplea la estrategia del ítem virtual desarrollada por Fischer, pero en este caso desdoblando un ítem en diferentes items virtuales en función del patrón de respuestas en los items precedentes.

Otra línea de trabajo también encaminada a la evaluación del cambio es la representada por los Modelos Lineales y de Medidas Repetidas para los parámetros de aptitud de los sujetos (Linear and Repeated Measures Models for the Person Parameters, Hoijtink, 1995). Estos modelos pretenden estimar el cambio entre ocasiones, al igual que el modelo **LLMT** anteriormente expuesto, o entre grupos de sujetos en una misma ocasión (p.e. hombres frente a mujeres). No obstante, la lógica que emplean es diferente a la expuesta para el **LLMT**: en lugar de modelar el cambio en el parámetro de dificultad del ítem, desdoblando el ítem en diferentes items virtuales, el cambio se modela en el parámetro de los sujetos.

Los Modelos Logísticos de Distribución Mixta (Mixture Distribution Rasch Models, Rost y Von Davier, 1995; Rost, 1997) responden a otra necesidad: la necesidad de aplicar el modelo de Rasch en el caso en que se sospeche que los datos no proceden de una única población sino de diferentes poblaciones. La utilidad del modelo estriba en que esas poblaciones no han de ser especificadas por el investigador, sino que es el modelo el que permite su identificación y la estimación de los parámetros correspondientes.

Por último, el Modelo Logístico de un Parámetro (*One Parameter Logistic Model*, **OPLM**, Verhelst y Glas, 1995) representa una extensión del modelo de Rasch sustancialmente diferente de las anteriores. La aportación de este modelo reside en la posibilidad de incorporar items con diferentes índices de discriminación. La estrategia que emplea se basa en introducir esos diferentes índices de discriminación como valores conocidos, en lugar de ser estimados por el modelo, como ocurre en el modelo de Birnbaum. Esta estrategia permite seguir considerando la puntuación total de cada sujeto como un estimador suficiente de la aptitud, con lo que el modelo conserva todas las propiedades que caracterizan al modelo de Rasch.

El siguiente gran apartado en que pueden clasificarse los modelos derivados del modelo de Rasch surge, como ya indicamos, de la adaptación de avances realizados sobre los modelos logísticos de 2 y 3 parámetros. En consecuencia, presentaremos primero cuáles han sido esos avances y luego continuaremos con las adaptaciones realizadas sobre el modelo de Rasch.

Extensiones de los modelos de Birnbaum.

Una de los avances más importantes desarrollados en el ámbito de la TRI fue la extensión de algunos de sus modelos a items con formato de respuesta politómico. Los primeros avances en este sentido se realizaron sobre el modelo logística de dos parámetros de Birnbaum y se deben a Samejima. Esta autora presentó el primer modelo logística para items politómicos: el Modelo de Respuesta Graduada, (**MRG)**, (Samejima,1969). Este modelo respondía a la necesidad de modelar un tipo de respuesta más sensible que la habitual hasta entonces, que se limitaba a un proceso de todo (acierto) o nada (fallo). Con ello se abría la posibilidad de aplicar los avances de esta teoría a escalas en que las diferentes respuestas de los sujetos a un ítem indicaban diferentes localizaciones de los sujetos en la dimensión latente, como sucede con las escalas tipo Likert, o con determinados items de rendimiento.

La estrategia que permitió a Samejima (1969) la aplicación de este modelo a items politómicos consiste en dividir la variable de respuesta politómica en una serie de variables dicotómicas y en especificar una función característica para cada una de ellas (curva característica de la categoría, **CCC**). En concreto, Samejima utilizó un procedimiento acumulativo, en el que la curva característica de la categoría "k" indica la probabilidad de alcanzar esa categoría o las siguientes, condicionada a la localización del sujeto en el rasgo (***P(Xi≥ k*** $|\theta$)). Este tipo de dicotomización permite estimar la probabilidad condicionada de que un sujeto alcance una determinada puntuación "***k***" a partir de la diferencia (***P(Xi ≥ k*** $|\theta$)- (***P(Xi≥k+1*** $|\theta$). Este es el motivo por el que Thissen y Steinberg (1986), autores de una de las clasificaciones más conocidas de los modelos de la TRI, denominan a los modelos que emplean esta estrategia "modelos diferenciales". Otra clasificación importante es la realizada por Mellenberg (1995), quien agrupa estos modelos bajo la denominación de *"respuesta acumulativa"*.

De la aplicación de esta estrategia al modelo de Rasch surgieron dos nuevos modelos, el Modelo de Escala de Clasificación (Rating Scale Model, MEC, Andrich, 1978, 1982: Andersen, 1977, Andersen, 1997) y el Modelo de Crédito Parcial (Partial Credit Model, MCP; Masters, 1982, Masters y Wright, 1997).

El Modelo de Crédito Parcial difiere del Modelo de Respuesta Graduada en que asume la igualdad de los parámetros de discriminación de cada categoría de respuesta del ítem y en que la división de las categorías de respuesta sigue el método de categorías adyacentes según la clasificación de Mellenberg (1995) o división por el total según la clasificación de Thissen y Steinberg (1986). Ello implica que las **CCC** indican la probabilidad de alcanzar la categoría k condicionada al grupo de sujetos que alcanzan esa categoría o la categoría anterior, k-l, dada una localización en el rasgo (***P(Xi = k*** $|\theta$***; Xi =k o Xi =k-l)***) Ello conlleva que los parámetros de localización de cada categoría no pueden ser interpretados como la dificultad de esa categoría, porque también incluyen información acerca de la categoría anterior.

Por su parte, el Modelo de Escala de Clasificación (**MEC**) puede ser considerado una caso particular del Modelo de Crédito Parcial, cuya particularidad estriba en que asume que las diferentes categorías de respuesta del ítem se encuentran equidistantes, es decir, que el incremento en la cantidad de rasgo que se requiere para pasar de una categoría a la siguiente es constante a lo largo de todas las categorías. En términos formales este supuesto implica que solo es necesario incluir un parámetro de dificultad por cada ítem. La dificultad de cada categoría vendrá dada por la combinación aditiva del parámetro del ítem y el incremento en la dificultad correspondiente al número de categorías acumuladas por debajo de una categoría dada. Este modelo ha sido aplicado fundamentalmente a datos procedentes de escalas con formato de respuesta tipo Likert (Andrich, 1978, 1982: Andersen, 1977, Andersen, 1997).

Estos dos modelos también han sido objeto de desarrollo posterior. En concreto, el Modelo Lineal de Escala de Clasificación (Lineal Rating Scale Model, Fischer y Parzer, 1991; Fischer, 1997) y el Modelo Lineal de Crédito Parcial (Lineal Partial Credit Model, Fischer y Povonocy,1994; Fischer, 1997) presentan extensiones de estos modelos susceptibles de descomponer el parámetro de dificultad en sus componentes cognitivas básicas, y de ser aplicados en dos o más ocasiones para medir el cambio de los sujetos. La estrategia empleada en estos modelos es análoga a la presentada en el modelo **LLMT** (Fischer, 1995).

Otras formulaciones alternativas del Modelo de Crédito Parcial son las presentadas por el Modelo de Etapas de Crédito Parcial (Steps Model to analize Partial Credit, Verhelst, Glas y Vries, 1997) y por el Modelo Secuencial (Sequential Model for Ordered Responses, Tutz, 1990, 1997). Ambos modelos tienen como objetivo eliminar la dependencia del parámetro de dificultad de cada **CCC** de la categoría anterior, que es el problema que presenta el Modelo de Crédito Parcial. Aunque formulados de forma independiente, ambos modelos son formalmente idénticos (Van der Linden y Hambleton, 1997). En ambos casos el método de división de las categorías de respuesta es el de respuesta continua (Mellenberg, 1995). Ello implica que las **CCC** indican la probabilidad de alcanzar la categoría **k+l** o siguientes condicionada al grupo de sujetos que alcanzan al menos la categoría **k**, dado un determinado nivel en el rasgo ($P(Xi \geq k+1 \mid \theta; Xi \geq k)$).

Una última generalización del MCP en el marco de los modelos de un parámetro es el denominado Modelo de Partición Ordenada (Ordered Partition Model, Wilson, 1992). Este modelo es aplicable al caso en que las diferentes categorías del ítem no pueden ser completamente ordenadas porque existen diferentes formas de alcanzar la misma puntuación en el ítem, y el investigador desea mantener esa distinción en el modelado de las respuestas de los sujetos.

En cuanto a las extensiones del modelo de dos parámetros, hay que señalar que se han desarrollado otras extensiones además del modelo de Respuesta Graduada. Los modelos de Crédito Parcial Generalizado y de Escala de Clasificación Generalizada representan generalizaciones de los modelos originales de un parámetro al modelo de dos parámetros (Muraki, 1992). La diferencia entre ambos es la misma que aparece en los modelos originales: el modelo de Crédito Parcial Generalizado introduce un parámetro de discriminación para cada **CCC**, mientras que el modelo de Escala de Clasificación Generalizada solo introduce un parámetro de discriminación por cada ítem.

Otras extensiones.

Como ha podido apreciarse tras la exposición de los modelos anteriores, el ámbito que es posible abarcar mediante modelos de la TRI es muy diverso. Y sin embargo, lo hasta aquí expuesto es sólo parte de lo que se ha avanzado desde la década de los 60. En concreto, la parte correspondiente a los modelos dicotómicos y politómicos. Algunas de las líneas de desarrollo restantes amplían la aplicabilidad de esta teoría:

Al ámbito de los tests con formato de respuesta de elección múltiple, especificando diferentes **CCC** para las diferentes alternativas de respuesta, a fin de extraer más información que la que proporciona el simple acierto/fallo (Bock, 1972; Thissen y Stenberg, 1984).

Al ámbito de los tests con formato de respuesta continua (Samejima, 1972).

Al ámbito de los tests de intentos múltiples (Spray, 1997), de velocidad (Roskam, 1997) y de potencia y velocidad, (Verhelst, Versralen y Jansen, 1997).

Al ámbito multidimensional (McDonald, 1967,1997; Reckase, 1985,1997).

Al ámbito de items no monótonos, a fin de poder modelar datos de proximidad (Andrich, 1997; Hoijtink, 1997).

Al ámbito no paramétrico, a fin de poder modelar datos provenientes de escalas ordinales (Mokken, 1971, 1997; Ramsay, 1991, 1997; Molenaar, 1997).

La simple observación de las fechas en que se han ido publicando los modelos presentados muestra claramente que la TRI sigue su expansión generando modelos cada vez más realistas, más flexibles, mejores en una palabra. Una buena muestra de ello se encuentra en publicaciones monográficas como las que le ha dedicado la revista *Applied Psychological Measurement* en números recientes (1995, Vol. 19, Nºl, 1996, Vol.20, Nº4) a los modelos politómicos y multidimensionales respectivamente; las revisiones sobre modelos multidimensionales de Reckase (1997), sobre modelos no paramétricos de Sijtsma (1998), y las compilaciones con aportaciones de los propios autores que presentan Fischer y Molenaar (1995), y Van der Linden y Hambleton (1997).

En resumen, en los últimos 30 años la proliferación de modelos es de tal magnitud que no existe ninguna clasificación que permita abarcarlos a todos y dar cuenta de sus semejanzas y diferencias. Hambleton (1989) ha llegado a decir que no hay límite para la cantidad de modelos que se pueden generar desde la TRI; por ello, señalaremos las tipologías más significativas y algunos intentos realizados para unificar este campo.

La clasificación ofrecida por Traub y Lain (1985) permite enmarcar a la TRI en el conjunto más general de los modelos de estructura latente. Los modelos de estructura latente son aquéllos cuyas variables teóricas no son observables directamente (variables latentes), sino a través de indicadores o variables manifiestas. Hay dos tipos de modelos, en base a la naturaleza de la variable latente, los modelos de clase latente y los modelos de rasgo latente.

En los primeros, la variable latente es discreta, bien nominal u ordinal, dando lugar a modelos de clase latentes categoriales y a modelos de clases latentes ordenadas.

En los segundos, la variable latente es continua y comprende los modelos de la TRI, los modelos del análisis factorial y los modelos de ecuaciones estructurales. Una presentación de los modelos de la TRI desde este punto de vista ha sido presentada por Heinnen (1993, 1996) y Rost y Langeheine (1997). En este contexto hay que precisar que si en la Teoría de los Tests se han utilizado mucho más los modelos de rasgo latente, también se han empleado los modelos de clases latentes en los tests de maestría o de competencia mínima (Dayton, 1991; Macready y Dayton, 1977, 1980).

Recientemente se han propuesto modelos mixtos resultantes de combinar un modelo de la TRI con un modelo de clase latente, como por ejemplo el modelo HYBRID de Yamamoto y Gitomer (1993).

Una clasificación bastante popular es la ofrecida por Bejar (1983a), dada su sencillez y claridad, aunque ya han pasado muchas cosas desde entonces. Los criterios empleados son el tipo de respuesta, la función matemática y la estructura de parámetros. El primer criterio considera el formato de respuesta de los items y da lugar a los modelos dicotómicos, los modelos politómicos, bien de respuesta nominal o de respuesta graduada, y los modelos continuos. El segundo criterio atiende al tipo de función matemática utilizada para modelar la probabilidad de respuesta, diferenciando entre modelos normales y modelos logísticos. El tercer criterio alude a la cantidad de parámetros que contiene la función de probabilidad, habiendo modelos de uno, dos, tres y cuatro parámetros, según consideren la

dificultad, la discriminación, el pseudo-azar y el descuido, respectivamente. Una relación de los modelos más representativos de cada categoría se encuentra en el cuadro 1.1.

La clasificación de Thissen y Steinberg (1986) relaciona los modelos según la estructura interna utilizada para modelar la probabilidad. Distinguen cinco tipos, denominados modelos binarios (Lord, 1952; Rasch, 1960; Birnbaum, 1968), modelos de diferencia (Samejima, 1969), modelos de división por total (Andrich, 1978; Bock, 1972; Masters, 1982), modelos de cola izquierda (Birnbaum, 1968) y modelos de cola izquierda y división por total (Samejima, 1972; Sympson, 1983; Thissen y Steinberg, 1984). Esta clasificación no ha tenido mucha aceptación.

Navas (1997) propone una sencilla clasificación de los modelos más conocidos, que puede servir para entender la variedad de formulaciones diferentes que existen dentro del marco de los modelos de TRI. Son tres las dimensiones o ejes en los que se basa esta clasificación. Primero, el tipo de respuesta que se obtiene al aplicar los tests: dicotómica, politómica o continua. Segundo, el tipo de función utilizada para relacionar la actuación del sujeto en el test con su nivel en la característica evaluada por el mismo. Si esta función matemática es la normal, se habla de modelos de ojiva normal; si es la logística, de modelos de ojiva logística. Tercero, el número de parámetros del ítem que, junto con el nivel del sujeto en la característica evaluada, influye de forma determinante en la actuación de éste en el test. Así, se dispone de modelos de uno, dos y tres parámetros, según se considere la dificultad (b), discriminación (a) y pseudoadivinación (c) como parámetros del ítem en el modelo (véase el cuadro 1.2).

Tipo de Respuesta	**Autor**
RESPUESTA DICOTÓMICA:	
Modelo logístico de 1 parámetro	Rasch (1960)
Modelo logístico de 2 y 3 parámetros	Birnbaum (1968)
Modelo logístico de 4 parámetros	McDonald (1967)
Modelos normales	Lord (1952,1953 a y b)
RESPUESTA POLITÓMICA:	
Nominal:	
Modelo de respuesta nominal	Bock (1972)
Modelo nominal modificado	Samejima (1979)
Modelo de elección múltiple	Thissen y Steinberg (1984)
Ordenada:	
Modelo de respuesta graduada	Samejima (1969)
Modelo de escala estimación	Andrich (1978)
Modelo de crédito parcial	Masters (1982)
Modelo de crédito parcial generalizado	Muraki (1992)
Modelo secuencial	Tutz (1990)
RESPUESTA CONTINUA:	
Modelo de respuesta continua	Samejima (1972)
Modelo continuo de Rasch	Müller (1987)

Cuadro 1.1.- Clasificación de modelos según los criterios de Bejar (1983a). Tomado de Hontangas (1997).

Como ya hemos señalado, las limitaciones de los modelos dicotómicos han conducido a que se preste cada vez mayor atención a los modelos politómicos y a los modelos multidimensionales. Las demandas de muchas situaciones de medida son atendidas mejor por estos modelos. Fruto de este interés ha sido el desarrollo de un número creciente de modelos y de publicaciones sobre los avances y sus problemas.

Los ***modelos politómicos*** suelen ser clasificados en función de la naturaleza ordinal o nominal de las alternativas de respuestas del ítem. Una distinción más refinada es ofrecida en los trabajos de Molenaar (1983), Mellenbergh (1995, 1996) y Engelenburg (1997). Estos autores consideran que un ítem politómico con k categorías puede verse como una serie de k-l items dicotómicos y clasifican los modelos por el procedimiento empleado en la segmentación de las categorías. Esta forma de proceder permite conocer la semejanza entre modelos que aparentemente pueden parecer muy distintos. Los procedimientos para efectuar la segmentación pueden ser nominal, acumulado, adyacente y contínuo. En la segmentación nominal, las dicotomías se forman combinando una categoría cualquiera, que sirve de referencia, con cada una de las demás (Bock, 1972). La segmentación acumulativa se construye con particiones que agrupan categorías bajas y altas respectivamente (Samejima, 1969; Molenaar, 1982). La segmentación adyacente se forma agrupando pares sucesivos de dos categorías (Masters, 1982; Muraki, 1992; Thissen y Steinberg, 1986). Y la segmentación continua se construye comparando cada categoría con la combinación de categorías superiores (Tutz, 1990).

Tipo de Respuesta	**Autor**
RESPUESTA DICOTÓMICA	
OJIVA NORMAL Modelos de 1, 2 y 3 parámetros	Lord (1952,1953a,1953b)
OJIVA LOGÍSTICA Modelos de1, 2 y 3 parámetros	Birnbaum (1957,1958a,1958b,1968); Rasch, 1960; Lord, 1980;
RESPUESTA POLITÓMICA	
Modelo de respuesta graduada	Samejima (1969, 1995,1997)
Modelo de respuesta nominal	Bock (1972, 1997)
Modelo de escala graduada	Andrich (1978); Andersen, 1997)
Modelo de crédito parcial	Masters, 1982; Masters y Wright, 1997; Muraki (1997); Verhelst, Glas y de Vries, (1997).
Modelo de elección múltiple	Thissen y Steinberg (1986,1997)
Modelo secuencial	Tutz (1990, 1997)
Modelo politómico de Rasch	Andersen, (1995); Fischer, (1995); Glas y Verhelst,
RESPUESTA CONTINUA	
Modelo de Samejima	Samejima (1972, 1973), (1995).
Modelo de Rasch	Müeller, (1987)
Modelo de Mellenbergh Ferrando (1995)[4]	Mellenbergh (1993,1994)

Cuadro 1.2.- Clasificación general de modelos TRI (Navas, 1997).

[4] Parte del modelo para items (tests) congenéricos de Jöreskog (1971) basado en el modelo general del análisis factorial lineal, y desarrolla una metodología general para calibrar parámetros invariantes desde el modelo TRI para respuestas continuas, basado en el enfoque del análisis factorial.

Los ***modelos multidimensionales*** fueron introducidos inicialmente por Lord y Novick (1968), Reckase (1972), Mulaik (1972) y Samejima (1974). El criterio de clasificación más interesante es la forma de interacción entre los rasgos latentes para producir la respuesta. Los modelos se dividen en compensatorios y no compensatorios. En los primeros, los rasgos se combinan de manera aditiva y los sujetos pueden suplir su bajo nivel en un rasgo por el nivel en otros rasgos (Doody-Bogan y Yen, 1983, Hattie, 1981, McKinley y Reckase, 1983; Reckase, 1985), mientras que en los segundos, los rasgos se combinan de forma multiplicativa y los sujetos necesitan varias capacidades simultáneamente para acertar un ítem (Sympson, 1978; Whitely, 1980). También se pueden clasificar en base a otros criterios, como la forma de la función matemática. En este caso, tenemos modelos muldimensionales normales y modelos multidimensionales logísticos. Entre los primeros se encuentran los modelos de Bock, Gibbons y Muraki (1985), Carlson (1987) o McDonald (1997) y entre los segundos los modelos de McKinley y Reckase (1983), Doody-Bogan y Yen (1983) o Sympson (1978). Una revisión de la historia, características y futuro de los modelos multidimensionales puede encontrarse en Reckase (1997). En castellano, se puede consultar el trabajo de Maydeu (1996).

Van der Linden y Hambleton (1997) presentan los modelos más recientes agrupados en 6 categorías y omiten los modelos para items dicotómicos por ser muy conocidos. La primera categoría aborda los modelos apropiados para items con formato de respuestas politómicas discretas ordenadas y no ordenadas. La segunda categoría comprende modelos para items en los que se considera el tiempo de respuesta y modelos en los que se hacen repetidos intentos en los items. La tercera categoría recoge los modelos multidimensionales desarrollados para medir varios componentes cognitivos o habilidades múltiples simultáneamente. La cuarta categoría se refiere a modelos no paramétricos. Se trata de modelos en los que se relajan los supuestos de manera que la forma de la función puede ser definida sin necesidad de especificar parámetros. La quinta categoría recoge modelos para items no monótonos que se caracterizan por datos de proximidad. Y la última categoría agrupa modelos que tienen supuestos especiales, tales como modelos mixtos con diferentes distribuciones, modelos basados en la dependencia local de los items o modelos para múltiples grupos simultáneos. Los modelos incluidos en cada una de las categorías se presentan en el cuadro 1.3.

Algunos autores han intentado unificar esta diversidad de modelos. Mellenbergh (1994b) ha propuesto una Teoría de Respuesta a los Ítems Lineal Generalizada (GLIRT) que presenta los modelos de la TRI integrados junto con otros modelos estadísticos en el marco más global del modelo lineal generalizado. En el mismo sentido se podrían situar los trabajos de Goldstein y Wood (1989) sobre el Modelo de Respuesta al ítem Lineal General (GLIRM) y el de McDonald (1982), que propone una ordenación de los modelos dentro del modelo lineal, así como los de McDonald (1986, 1989) que intentan situar tanto a la TRI como a la TCT en el marco de los modelos estadísticos multivariantes. A un nivel menos general, también cabe destacar el trabajo realizado por Ferrando (1996) sobre las relaciones entre la TRI y el análisis factorial.

Tipo de Modelo	**Autor**
MODELOS POLITÓMICOS	
Modelo de respuesta nominal	Bock
Modelo de elección múltiple	Thissen y Steinberg
Modelo de escala de estimación	Andersen
Modelo de respuesta graduada	Samejima
Modelo de crédito parcial	Masters y Wright
Modelo de etapas	Verhelst, Glas y de Vries
Modelo secuencias ordenado	Tutz
Modelo de crédito parcial generalizado	Muraki
MODELOS PARA TIEMPO DE RESPUESTA O INTENTOS MÚLTIPLES	Verheist, Verstralesn y Jansen
Modelo de tiempo límite	Roskam
Modelo tiempo limite y velocidad	Spray
Modelo de intentos múltiples	
MODELOS DE COMPONENTES COGNITIVOS O HABILIDADES MÚLTIPLES	
Modelo logística lineal	Fischer
Modelo con predictores observados	Zwinderman
Modelo multidimensional normal	McDonald
Modelo multidimensional logística lineal	Reckase
Modelo multidimensional loglineal	Keldennan
Modelo multicomponentes	Embretson
Modelo multidimensional logística lineal	Fischer y Seliger
MODELOS NO PARAMÉTRICOS	
Modelo de Mokken dicotómico	Mokken
Modelo de Mokken politómico	Molenaar
Modelo de análisis funcional	Ramsay
Modelo de coseno hiperbólico	Andrich
Modelo de paralelogramo	Hoijtink
MODELOS CON SUPUESTOS ESPECIALES	
Modelo de grupos múltiples	Bock y Zimowski
Modelo de logística mixto	Rost
Modelo de conjuntiva	Jannarone
Modelo de desajustes	Hutchinson

Cuadro 1.2.- Modelos según la clasificación de van der Linden y Hambleton (1997). Tomada de Hontanga (1997).

Desde un punto de vista aplicado, los modelos más utilizados son los logísticos unidimensionales con uno, dos y tres parámetros para items dicotómicos. Estos modelos se han empleado principalmente para medir inteligencia, aptitudes y rendimiento y se han hecho muy pocos intentos en áreas relativas a la medida de la personalidad, actitudes e intereses u otros rasgos similares de la conducta típica de los sujetos. Actualmente, los modelos politómicos están empezando a dar el salto del ámbito de la investigación al mundo aplicado, sin embargo, los multidimensionales permanecen todavía dando vueltas en los ordenadores de los equipos de investigación más experimentados.

En definitiva, la TRI esta respondiendo a las necesidades de la sociedad, que demanda una medida de objetividad contrastada y contrastable. Como también esta respondiendo a las necesidades de la comunidad científica, aportando instrumentos capaces de responder a demandas cada vez más sofisticadas, como la de modelar el funcionamiento cognitivo de los sujetos. De hecho, el binomio TRI-Psicología Cognitiva se esta revelando como

una de las líneas de desarrollo mas prometedoras como veremos, con un poco más de detalle, en el apartado dedicado a las perspectivas de futuro de esta teoría.

Hasta aquí hemos expuesto los fundamentos de la TRI y los modelos básicos y sus desarrollos más relevantes. A continuación procederemos a describir resumidamente el procedimiento que suele seguirse en la práctica para la elección de estos modelos, la estimación de los parámetros y la comprobación del ajuste de los mismos a los datos empíricos.

1.5. Comprobación de los Modelos.

Como Señala Muñiz, (1997), *"Comprobar que el modelo de TRI elegido es el adecuado constituye por razones obvias un aspecto central en la aplicación de los modelos"* (p.96). Al conjunto de actividades que hemos de realizar para la elección de un modelo adecuado a las aplicaciones prácticas se denomina "comprobación del modelo". Ello supone que una vez definida adecuadamente la variable a medir, elaborados los items destinados para medirla y realizada una aplicación a un conjunto amplios de personas, hemos de comprobar en primer lugar si los datos cumplen con el supuesto de unidimensionalidad y, una vez salvado este primer obstáculo, hemos de decidir cuál de los modelos (1, 2 o 3 parámetros) es el más adecuado a nuestro datos; para ello se suele utilizar la información aportadas por los índices clásicos de los items (índices de discriminación, de dificultad). Elegido un modelo, se procede a la estimación de los parámetros del modelo elegido y al cálculo de las puntuaciones de los sujetos. Si el modelo elegido es el adecuado, éste debería ajustarse a los datos. Es por ello, que la estimación de los parámetros y el ajuste del modelo son dos aspectos fundamentales en este proceso de comprobación de los modelos. A continuación describiremos brevemente estos aspectos.

1.5.1. Estimación de parámetros

Los métodos de estimación de los parámetros de cualquier modelo TRI se basan fundamentalmente en el principio de máxima verosimilitud, en criterios bayesianos o en estrategias heurísticas. Esta estimación suele hacerse en dos situaciones diferentes, de un lado nos encontramos con situaciones en las que la estimación ha de hacerse conjuntamente y, de otro lado, la estimación se ha de realizar de manera condicional. En la primera se desconocen tanto los parámetros de items como los de los sujetos y ambos han de ser estimados simultáneamente a partir de las respuestas (v.gr., calibración de un banco de items). En la segunda se conocen los parámetros de los items, pero se desconocen los de los sujetos, estimándose éstos a partir de aquéllos y de las respuestas de los sujetos (v.gr., aplicación de un test adaptativo).

En primer lugar, el método de *máxima verosimilitud* busca los valores que hacen más probable la obtención de los datos empíricos a partir del modelo. El método de máxima verosimilitud condicional se emplea para estimar la habilidad de los sujetos, conocidos los valores de los parámetros (Lord, 1980). También se denomina así al método empleado para estimar los parámetros de los items en el modelo de Rasch, condicionando la función de verosimilitud sobre el número de respuestas correctas (Andersen, 1972; Rasch, 1960). En caso de que haya que estimar los parámetros de items y sujetos simultáneamente se emplean los métodos de máxima verosimilitud conjunta (Lord, 1974, Birnbaum, 1968) o de máxima verosimilitud marginal (Bock y Liebennan, 1970; Bock y Aitkin, 1981).

En segundo lugar, los métodos bayesianos se basan en combinar la función de verosimilitud de los datos muéstrales con una distribución adicional, que se supone siguen los parámetros (distribución a priori), dando lugar a

una distribución a posteriori. Las estimaciones de los parámetros son los valores que maximizan la distribución a posteriori, considerando como el máximo de la distribución la moda (Swaminathan y Guilfford, 1982, 1985, 1986) o la media (Bock y Mislevy, 1982).

En tercer lugar, los métodos heurísticos se basan en la equivalencia que existe entre algunos modelos de TRI y ciertos estadísticos de la TCT (Urry, 1974, 1976). En la actualidad, estos métodos se utilizan únicamene para obtener los valores iniciales de los estimadores en algunos programas informáticos. También existen otras posibilidades como el método de mínimos cuadrados ponderados (Baker, 1992), el método de chi-cuadrado mínimo (Berkson, 1955) o los métodos PROX y UCON en el contexto del modelo de Rasch (Wright y Stone, 1979).

Los métodos de estimación requieren el uso de algoritmos de aproximación numérica, dado que no se puede obtener directamente una solución resolviendo las ecuaciones resultantes de derivar la función de verosimilitud. Las técnicas utilizadas son el algoritmo de Newton-Raphson, basado en la expansión de las series de Tylor (Bunday, 1984), y el método de puntuación de Fisher (Rao, 1965).

Afortunadamente, hay una gran cantidad de programas para estimar los parámetros de casi todos los modelos, tales como ANCILLES, OGIVA (Urry, 1974, 1976); ASCAL (Vale y Gialluca, 1985); BICAL (Wrigth y Mead, 1976); BIGSTEPS (Linacre y Wright, 1997); BILOG (Mislevy y Bock, 1990); BILOG-MG (Zimowski, Muraki, Mislevy y Bock (1996); BIMAIN (Muraki, Mislevy y Bock, 1987; Zimowski, Muraki, Mislevy, y Bock, 1993); LOGIST (Wingersky, Barton y Lord, 1982); METRIX (Renom, 1992b); MIRTE (Carlson, 1987); MULTILOG (Thissen, 1991); NOHARM (Fraser, 1988); PARSCALE (Muraki y Bock, 1996); PML (Gustaffson, 1980); QUEST (Adams y Khoo, 1995); RASCAL, (ASC, 1988); RIDA (Glass, 1990); RSP (Glass y Ellis, 1993); TESTAT (Stenson y Wilkinson, 1986); WINMIRA (van Davier, Smith, y Makov, 1995) y XCALIBRE (ASC, 1988). También existen estudios que comparan la eficiencia de algunos de estos programas (Mislevy y Stocking, 1989; Yen, 1987; Vale y Gialluca, 1988) y su capacidad para recuperar parámetros de varios modelos mediante estudios de simulación (Harwell y Janosky, 1991; Hulin, Lissak y Drasgow, 1982; López-Pina, 1995; Reise y Yu, 1990; Stone, 1992).

Una presentación breve de los procedimientos de estimación más comunes puede encontrase en cualquiera de los manuales sobre la TRI (Lord, 1980; Hambleton y Swaminathan, 1985; Martínez-Arias, 1995) y también en las revisiones de Baker (1987) o Mislevy (1986). Ahora bien, queremos destacar las aportaciones de Baker (1992) por su rigor y exhaustividad, y la de López-Pina (1995) por su claridad y orientación didáctica.

1.5.2. Ajuste del modelo.

Las ventajas de este enfoque se obtienen cuando el modelo es apropiado y ajusta a los datos. La existencia de múltiples índices de ajuste, ninguno de ellos completamente satisfactorio, aconseja utilizar varios tipos de estrategias para recoger evidencias que permiten hacer un juicio global de la adecuación del modelo (Hambleton y Rogers, 1991). En este sentido, los aspectos que hay que analizar son el cumplimiento de los supuestos, el ajuste del modelo y las ventajas esperadas. En primer lugar, hay que comprobar el grado de cumplimiento de los supuestos de los que ya hemos hablado, como la unidimensionalidad y la independencia local, y también de aquellos que son específicos para cada modelo, como la ausencia de adivinación en el modelo de dos parámetros, la igualdad de discriminación en el modelo de un parámetro o la ausencia de velocidad. En segundo lugar, hay que estimar el ajuste entre los datos predichos por el modelo y los datos obtenidos en la muestra de estudio, mediante análisis de los

residuales, índices basados en chi-cuadrado o la comparación de las distribuciones de puntuaciones predichas y observadas. En tercer lugar, hay que comprobar que las ventajas atribuidas al modelo se producen, como la invarianza de los parámetros de los sujetos a través de diferentes grupos de items y la invarianza de los parámetros de los items en diferentes muestras de sujetos. Una descripción de las principales técnicas a utilizar para cada uno de estos aspectos se pueden encontrar en Hambleton y Swaminathan (1985), Hambleton, Swaminathan y Rogers (199l), Martínez-Arias (1995) y López-Pina e Hidalgo (1996).

Una vez que se ha comprobado que el modelo propuesto presenta un ajuste adecuado a los datos, las estimaciones de los parámetros de los sujetos producidas por el modelo ofrecerán la medida de cada sujeto en el rasgo latente de interés. La precisión de estas medida suele representarse mediante la curva característica del test. Veamos a continuación estos concepto.

1.6. Curva Característica del Test.

El concepto de curva característica del test (**CCT**) es similar al concepto de **CCI**. Su interés principal se centra en que la **CCT** sirve de nexo entre la TRI y la TCT posibilitando, entre otras, la interpretación de los resultados y la equiparación de las puntuaciones de los sujetos.

La curva característica del test es la suma de las curvas características de los items que componen el test, es decir, para obtener un determinado nivel de θ se suman los valores de $P(\theta)$ de cada ítem del test para ese nivel. Formalmente puede expresarse como:

$$CCT = \sum_{i=1}^{k} P_i(\theta) \tag{1.17}$$

siendo k el número de ítems.

Sus valores indican la relación que existe entre el nivel en el rasgo latente θ de un determinado sujeto y el patrón de respuesta esperado en el test.

Como ya hemos señalado, la TRI centra su interés en la estimación del valor de θ como valor del rasgo latente que se quiere medir; en este sentido, un test particular es un indicador de dicho rasgo y, por consiguiente, consta de dos componentes aditivos (la verdadera puntuación de θ en el test y el componente de error de medida correspondiente). Veamos como se pueden estimar dichos componentes.

1.6.1. Puntuación verdadera en el test.

La puntuación verdadera en el test de un sujeto al que se ha estimado una determinada puntuación $\theta=\theta_j$ mediante un determinado modelo de TRI viene dada por las **CCI** que componen el test, es decir, se estima como la suma de las probabilidades $P(\theta_j)$ de cada ítem:

$$V_j = \sum_{i=1}^{k} P_i(\theta_j) \tag{1.18}$$

donde **k** es el número de items y $P(\theta_j)$ el valor correspondiente a cada **CCI** para $\theta=\theta_j$.

Como puede observarse, el valor de la puntuación verdadera se corresponde con el valor generado por la curva característica del test para $\theta=\theta_j$.

1.6.2. Error típico de medida.

Para estimar el error de medida, la TRI propone dos estimaciones diferentes, una a nivel de cada items del test y otra a nivel de test:

A nivel de cada ítem el *error de medida* se calcula como la diferencia entre la puntuación empírica menos la puntuación verdadera. Formalmente,

$$e_i = 1 - P(x = 1 \mid \theta) \tag{1.19}$$

A nivel del test, para determinar el nivel real en que un sujeto posee la característica o rasgo latente que mide el test, se utiliza la siguiente formula general:

$$P\left(\hat{\theta} - E.\text{máx.} \leq \theta \leq \hat{\theta} + E.\text{máx.}\right) < \alpha \tag{1.20}$$

donde $E.máx = z_c\sigma_e$

El error típico de medida σ_e para un determinado nivel de $\theta = \theta_j$ puede calcularse como la raíz cuadrada de la varianza muestral de la estimación del parámetro θ ofrecida por el test:

$$\sigma_e = \sigma_{\hat{\theta}|\theta} = \sqrt{\sum_{i=1}^{k} P_i(\theta_j)Q(\theta_j)} \tag{1.21}$$

donde **k** es el número de items del test, $P_i(\theta_j)$ es el valor de las **CCI** para $\theta=\theta_j$ y $Q_i(\theta_j)$ es igual a 1- $P(\theta_j)$.

A su vez, la inversa de la varianza muestral de la estimación del parámetro θ ofrecida por el test nos da información de la precisión de la medida y recibe el nombre de función de información del test.

1.7. Función de información.

La función de información del test expresa la precisión de las estimaciones de las puntuaciones en el rasgo. Sustituye al coeficiente de fiabilidad de la Teoría Clásica de Tests, sobre el que presenta cuatro claras ventajas. En primer lugar, la función de información expresa la precisión del test como la combinación aditiva de la precisión de cada ítem. En segundo lugar, puede ser, obtenida a priori, antes de ser aplicado el test a la población de interés. Unicamente se requiere conocer los parámetros de los items de que se compone el test, y el rango de valores en ese rasgo que caracterice a esa población. En tercer lugar, esa función de información indica la precisión del test para cada puntuación estimada. Y en cuarto lugar, esa estimación de la precisión es independiente de la población. Veamos como se llega a la formulación de la función de información del test. Para ello comenzaremos con la presentación de otra función de información, la de la puntuación observada en un test (X). A diferencia de la anterior, esta función indica la cantidad de información acerca del rasgo que ofrecen las puntuaciones observadas en el test.

La función de información para una puntuación X es, por definición, la razón entre el cuadrado de la pendiente de la regresión de X sobre θ y el cuadrado del error típico de medida de X para un θ dado (Birnbaum, 1968)

$$I(\theta, X) = \frac{\frac{\partial}{\partial\theta}(\mu_{x|\theta})^2}{\sigma^2_{x|\theta}} \qquad (1.22)$$

donde el numerador es la derivada primera con respecto a θ de la curva de regresión, o curva de medias de X condicionadas a θ. Este valor expresa la pendiente de la curva de regresión en el punto θ. El denominador expresa la varianza de la puntuación X alrededor de la curva de regresión o media condicionada, es decir, la varianza del error de medida en ese punto. Esta expresión deja claro que la información crece a medida que la pendiente de la curva aumenta y a medida que la varianza de error disminuye.

La Teoría de Respuesta a los Items no sólo proporciona la puntuación observada en el test. También proporciona otra puntuación entre cuyas propiedades figura la de invarianza respecto del conjunto de items administrados. Se trata de la puntuación estimada en el rasgo . Esta puntuación es el estimador de máxima verosimilitud de la aptitud θ. Birnbaum (1968) también derivó la función de información correspondiente a esta puntuación aplicando la misma fórmula empleada en la función de información de las puntuaciones observadas. En este caso son necesarias las formulas asintóticas para la regresión y para la varianza:

$$I(\theta, \bar{\theta}) = \frac{\frac{\partial}{\partial\theta}(\mu_{\bar{\theta}|\theta})^2}{\sigma^2_{\bar{\theta}|\theta}} \qquad (1.23)$$

El resultado de operar con el numerador es la pendiente de la curva de regresión en el punto θ. Dado que lo que se considera es la regresión asintótica y en ese caso el valor esperado de la media del estimador de la aptitud condicionado a un determinado nivel de aptitud es ése nivel de aptitud, , y la pendiente en ese punto tiende a la unidad.

En cuanto al denominador, la estimación de la varianza de error, es decir, la varianza de la distribución de las estimaciones de la aptitud condicionadas a un determinado valor en el rasgo θ, puede resolverse acudiendo a las propiedades de las estimaciones máximo-verosímiles. Bajo ciertas condiciones (ver Santisteban, 1990, p.280) la varianza de la distribución de los estimadores de máxima verosimilitud de un parámetro θ responde a la siguiente expresión:

$$Var(\hat{\theta} \mid \theta) = \frac{1}{E\left(\frac{\partial \ln L}{\partial\theta}\right)^2} \qquad (1.24)$$

donde L es la función de verosimilitud, y $ln\ L$ su logaritmo neperiano. Esta expresión puede transformarse bajo ciertas circunstancias (Santisteban, 1990, p.299) en la siguiente expresion:

$$Var(\hat{\theta} \mid \theta) = \frac{1}{\sum_{i=1}^{k} \frac{P'^{2}_{i}(\theta)}{P_i(\theta)Q_i(\theta)}} \qquad (1.25)$$

Siendo P'_i la derivada primera de $P_i(\theta)$, y siendo $Q_i(\theta)=1-P_i(\theta)$.

En definitiva, sustituyendo y considerando que el numerador tiende a la unidad, la función de información del test puede expresarse como:

$$I(\theta) = I(\theta \mid \hat{\theta}) = \sum_{i=1}^{k} \frac{P'^{2}_{i}(\theta)}{P_i(\theta)Q_i(\theta)} \qquad (1.26)$$

Esta expresión muestra como la función de información del test puede ser caracterizada, bajo ciertas condiciones, como la inversa de la varianza muestral de la estimación del parámetro θ ofrecida por el test (Birnbaum, 1968).

Por otra parte, la función de información del ítem viene dada por la siguiente expresión (Birnbaum, 1968):

$$I_i(\theta) = \frac{P'^{2}_{i}(\theta)}{P_i(\theta)Q_i(\theta)} \qquad (1.27)$$

Esta expresión indica que la información de un ítem condicionada a una determinada puntuación en el rasgo es directamente proporcional a la pendiente de la **CCI** (regresión ítem-rasgo) en ese punto, e inversamente proporcional a la varianza muestral de la estimación en ese punto. Con esta definición de función de información del ítem llegamos a la conclusión de que la función de información del test puede obtenerse simplemente sumando las funciones de información de los items, tal como queda recogido en dicha expresión.

La función de información de las puntuaciones observadas en el test, $I(\theta,X)$, y la función de información de las puntuaciones estimadas en el rasgo o función de información del test, **,** están relacionadas. La relación es la siguiente: $I(\theta,X) \leq$, es decir, proporciona la cota máxima de información a que puede llegar $I(\theta,X)$. La forma en que $I(\theta,X)$ puede llegar a ofrecer esa cantidad de información o precisión consiste en aplicar las ponderaciones adecuadas a las puntuaciones de los items contenidos en el test. Estas ponderaciones dan más peso a los items más informativos, esto es, aquellos items que, entre otras características, son los que presentan mayor poder discriminante. Información detallada acerca de los valores en el rasgo en que las funciones de información de los items alcanzan su máximo, de la cantidad de información obtenida en ese punto y de las ponderaciones que hay que aplicar a las puntuaciones en los items, para que la puntuación total conlleve la cantidad máxima de información, puede encontrarse en Lord y Novick (1968), Lord (1980), Hambleton y Swaminathan (1985), Crocker y Algina (1986), entre otros, y en castellano en Santisteban (1990).

Una de las aplicaciones mas útiles de la función de información es la comparación de la eficiencia con que distintos tests miden el mismo rasgo, o la comparación de la eficiencia con que diferentes métodos de puntuación

aplicados a un mismo test miden el rasgo. Estas aplicaciones se llevan a cabo mediante el cociente denominado "eficiencia relativa", que en el caso de la comparación de dos tests X e Y presenta la siguiente expresión:

$$ER(Y, X) = \frac{I(\theta, Y)}{I(\theta, X)} \tag{1.28}$$

La eficiencia relativa de dos tests varía en función del nivel en la aptitud, y muestra en cada caso cuál es el test que ofrece una medida mas precisa.

Para concluir, presentaré algunas consideraciones acerca de la relevancia del concepto de función de información. Esta nueva aproximación al problema de la precisión de las medidas ofrecidas por los tests acaba con tres de los problemas centrales de la Teoría Clásica de Tests: el problema de la dependencia de la población, el problema de la homocedasticidad, y el problema de la estandarización. Respecto del primer problema, la función de información, al igual que la **CCI,** es una función condicionada a cada valor en el rasgo (ambas son curvas de regresión), y por tanto, es independiente de la distribución del rasgo en la población. Respecto del segundo problema, la función de información no sólo reconoce, sino que modeliza las variaciones en el grado de precisión de las medidas de sujetos con diferentes localizaciones en el rasgo. Y en cuanto al tercer problema, la función de información permite adaptar los tests a las características de las poblaciones de interés.

En definita la TRI presenta, en general, grandes ventajas respecto al modelo clásico. A continuación vamos a describir brevemente la ventajas y, al mismo tiempo, los inconvenientes de esta perspectiva.

1.8. Ventajas y limitaciones.

Las ventajas más destacables de la TRI son las referidas a la invarianza de parámetros, el tratamiento dado al error de medida, el estatus científico de la mísma, y la interpretación de las puntuaciones y sus aplicaciones prácticas (Fisher y Molenaar, 1995; Hambleton y Swaminathan, 1985; Martínez-Arias, 1995; Muñiz, 1997b; Lord, 1980; Rasch, 1960):

La invarianza de parámetros posibilita que, si el modelo ajusta a los datos, las características del test no dependan de la muestra en la que es analizado y las medidas de los sujetos no dependan del test utilizado.

El error de medida varía en los diferentes niveles del rasgo, pudiéndose conocer dónde se mide con mayor o menor precisión mediante la función de información que, además, permite evaluar la contribución individual de los items.

El mayor estatus científico procede de la posibilidad de contrastar empíricamente los supuestos en los que se sustenta, y también del nivel de medida alcanzado que permite construir escalas de intervalo.

La versatilidad para interpretar los resultados hace que se pueda aplicar tanto a tests normativos como a tests referidos al criterio. Puesto que la habilidad de los sujetos y la dificultad de los items están en la misma escala, es posible dar una puntuación en esa escala, describir lo que son capaces de hacer y ofrecer información diagnostica.

Las aplicaciones prácticas comprenden numerosos aspectos, entre los que destacan la elaboración de bancos de items, la construcción de tests con propiedades conocidas, la generación automática de items, los tests adaptativos informatizados, el análisis del funcionamiento diferencial de los items y la equiparación de puntuaciones.

Sin embargo, no todo son ventajas, también presenta limitaciones. Entre las más destacables señalaremos las siguientes:

El elevado tamaño muestral. Los modelos de la TRI requieren muestras muy grandes para que las estimaciones de los parámetros sean estables (Hambleton y Rogers, 1991). Y el problema es todavía mayor cuando se trabaja con items de formato de respuesta politómica, donde el número de parámetros a estimar se multiplica por el número de alternativas de respuesta. En casos en que sólo se disponga de muestras pequeñas lo más recomendable (Lord, 1983) es tratar de ajustar los modelos más parsimoniosos, que son aquellos que pertenecen a la familia de modelos de Rasch.

Los supuestos de los modelos son muy exigentes. La utilización de supuestos más fuertes restringe su aplicabilidad porque muchas veces no se cumplen o se cumplen sólo parcialmente (Lord y Stocking, 1988). Uno de los problemas más inquietantes es la unidimensionalidad, puesto que todos los items implican algún grado de multidimensionalidad. La combinación de rasgos adicionales con la dimensión principal en cantidades o proporciones desconocidas en los items dificulta la interpretación de los parámetros y, en general, de lo que se está midiendo. Esto puede afectar más a unas aplicaciones que otras, como aquéllas que usan diferentes items para medir a distintos sujetos.

La complejidad de la estimación de los parámetros de los modelos. Este problema afecta especialmente a los modelos multiparamétricos (2 y 3 parámetros) y a los modelos multidimensionales. Sin embargo, mientras los métodos de estimación de máxima verosimilitud marginal y los métodos bayesianos parecen haber reducido el problema en los modelos multiparamétricos, los modelos multidimensionales continúan presentando dificultades en este aspecto (Wright y Stone, 1979; Hambleton, 1994b; Maydeu, 1996).

Las técnicas existentes para comprobar el ajuste presentan múltiples inconvenientes con algunos modelos. La mayoría de modelos para ítems politómicos, cuando se aplican a tests con más de 5 ítems, no tienen índices de ajuste con distribución conocida, a excepción de aquellos que se derivan del modelo de Rasch. A lo sumo se puede contrastar qué modelo, entre varios alternativos, presenta mejor ajuste mediante pruebas basadas en razones de verosimilitud (Thissen y Steinberg, 1997). Ahora bien, mejor ajuste puede distar mucho de buen ajuste.

El estatus métrico de las puntuaciones que ofrecen estos modelos. Este punto sólo está resuelto para la familia de modelos de Rasch. El resto de modelos continúan asumiendo un nivel de medida (generalmente de intervalo) que no puede demostrarse (Muñiz y Hambleton, 1992).

En resumen, todas estas ventajas, en su conjunto, suponen una clara mejora respecto a los planteamientos del modelo clásico y son, sin lugar a dudas, las mejores razones para justificar el carácter hegemónico que actualmente tiene este modelo. La TRI se presenta como un marco ilimitado de referencia para resolver problemas de medición.