CIENCIA Y SOCIEDAD
Volumen XXIII, Número 3
Julio - Septiembre 1998

NUEVAS PERSPECTIVAS DE LA MEDICIÓN EN PSICOLOGÍA

Dr. José V. Díaz *

RESUMEN

Los intentos de hacer medición en Psicología, agrupados globalmente bajo la denominación de *Psicometría*, ha tomado históricamemte dos modalidades: una unida a la medición de estímulos físicos o psicológicos, que busca desarrollar procedimientos para construir *escalas*, (de ahí el nombre *Escalamiento*), capaces de estimar los valores de los estímulos, y que ha estado asociada a la Psicofísica y a la Psicología experimental (Weber, Fechner, Wundt, Thurstone), y la otra modalidad, la *Teoría de los Tests*, ha desarrollado teorías y técnicas que facilitan la construcción de instrumentos de medida psicológica (los tests), que han servido por años para asignar números a atributos o a comportamientos humanos. Esta segunda modalidad ha estado muy vinculada a la Psicología de las diferencias individuales, propugnada por Galton y demás autores de la escuela inglesa, principalmente Spearman.

PALABRAS CLAVE

Psicología, Psicometría, Tests, Escalas, Matriz de Datos.

La disciplina que hace posible la medición en Psicología, la Psicometría, debe ser visualizada, como un conjunto de modelos formales: que posibilita la medición de variables psicológicas, que se centra en las condiciones que hacen posible el proceso de medición y que establece las bases para que estos procesos se realicen de la forma adecuada. Dentro de la psicometría se han desarrollado varias teorías, como luego se verá, que proponen diversos modos de proceder para realizar la medición en psicología. Pero

(*) Profesor Titular de Psicometría de la Universitat de Valencia, España

todas ellas se basan en un mismo modelo, que denominaremos, psicométrico.

El modelo psicométrico pertenece al grupo de los modelos axiomáticos, que buscan definir los aspectos estructurales que subyacen detrás de los fenómenos. Estos modelos teóricos suelen ser expresados en lenguaje formal, cosa que facilita su comprensión y explicación. La calidad y aceptación de estos modelos teóricos depende del grado en que los datos encontrados empíricamente en la medición se ajustan a los datos teóricos generados por el modelo.

En todo modelo teórico hay que distinguir en primer lugar un conjunto de supuestos o hipótesis fundamentales y luego una serie de índices y derivaciones que permiten operar y explicar mejor los fenómenos. El modelo teórico sobre el cual se basa la Psicometría se fundamenta, pues, sobre estos supuestos:

1° Existe en los sujetos que van a ser medidos un *rasgo latente* o *atributo*, rasgo que suele ser denominado por **Y** en la Teoría Clásica (TCT) y por θ en la Teoría de la Respuesta al Item (TRI).

2° Luego se elige un conjunto de comportamientos observables capaces de medir este rasgo. Este conjunto total de items se le llama dominio del atributo (**X**): si se extrae del mismo una muestra representativa de **n** items se obtiene un test.

3° Aplicada esta muestra de items a los sujetos, estos dan respuestas diversas, dependiendo de su nivel de aptitud: respuestas que valoradas adecuadamente dan en su conjunto una puntuación del sujeto ($\mathbf{X}_i$), sobre las que se pueden hacer inferencias válidas acerca del nivel de aptitud de los sujetos.

Los datos anteriores suelen disponerse en una matriz, denominada matriz de datos, que presenta los siguientes elementos:

A. Un conjunto **n** de items, que pertenece al dominio del atributo, colocados en la fila primera.

B. Una muestra **N** de sujetos que ha sido extraída de la población, colocados en la primera columna.

C. Tres clases de puntuaciones:

   $\mathbf{x_{ji}}$: que es la respuesta del sujeto j al item i (puntuaciones elementales).

   $\mathbf{X_{j.}}$: que son las respuestas del sujeto j a los **n** items del test (puntuaciones del sujeto).

   $\mathbf{X_{.i}}$: que son las respuestas de todos los sujetos al item I (puntuaciones al item).

TABLA N° 1
Matriz de Datos

Item	1	2	3	...	i	...	n	$X_{j.}$
Sujetos 1	x_{11}	x_{12}	x_{13}	...	x_{1i}	...	x_{1n}	$X_{1.}$
2	x_{21}	x_{22}	x_{23}	....	x_{2i}	...	x_{2n}	$X_{2.}$
3	x_{31}	x_{32}	x_{33}	...	x_{3i}	...	x_{3n}	$X_{3.}$
...	...	...	...	...	...	...	...	...
j	x_{j1}	x_{j2}	x_{j3}	....	x_{ji}	...	x_{jn}	$X_{j.}$
....	...	...	...	...	...	...	...	...
N	x_{N1}	x_{N2}	x_{N3}	...	x_{Ni}	....	x_{Nn}	$X_{N.}$
$X_{.i}$	$X_{.1}$	$X_{.2}$	$X_{.3}$	...	$X_{.i}$	...	$X_{.n}$	

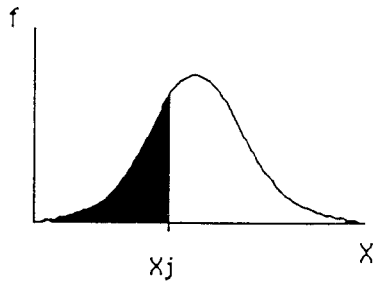
La selección de ítems de la muestra se hace utilizando las puntuaciones de los items (**X.i**), calculando una serie de índices o parámetros, que varían de acuerdo a la teoría de los tests adopta-

da. Así se adopta la Teoría Clásica de los Tests (TCT). Este cálculo se hace mediante los índices: de dificultad de los items (p_i), de discriminación interna (r_{ix}) y del índice de discriminación externa (x_{iy}) y el análisis de errores.

Atendiendo a las puntuaciones de los sujetos ($\mathbf{X}_j$) y a la distribución de frecuencias de las puntuaciones de los sujetos de la población, se pretende acomodar la misma a uno de los modelos propuestos. En la TCT suele utilizarse en exclusiva la curva de las probabilidades de Gauss, también llamada *Curva normal*, que está representada en la gráfica N° 1:

GRÁFICA N° 1

Curva normal de las frecuencias de las puntuaciones



En la Curva Normal pueden distinguirse dos elementos: la altura de la curva en un punto determinado ($\mathbf{X}_j$), y el area de la curva a partir de una puntuación concreta, cuyas ecuaciones son respectivamente:

$$\varphi_x = \frac{1}{2\pi\sqrt{\sigma}}\, e^{-\frac{(X-\mu)^2}{2\sigma^2}} \qquad \phi_X = \int_{-\infty}^{\alpha} \frac{1}{2\pi\sqrt{\sigma}}\, e^{-\frac{(X-\mu)^2}{2\sigma^2}}\, d_x$$

Las tablas de la Curva Normal que aparecen en los manuales de Psicometría permiten calcular el valor de estos elementos una vez conocidas las puntuaciones tipicas Z de los sujetos.

$$Z = \frac{X_j - \overline{X}_j}{s_x}$$

La TRI, además de la curva normal, utiliza también la logística, que expresa la probabilidad de obtener dichas puntuaciones, que por cierto resulta mas fácil de tratar desde el punto de vista matemático.

Teorias de los tests

Las aportaciones del Binet y Simon, (1905,1908,1911) resultaron sumamente importantes para el desarrollo de la Teoría de los Tests, ya que centraron la medición de la inteligencia en la evaluación de las *funciones mentales*, descartando las pruebas senso-motoras y antropométricas utilizadas en los primeros conatos de medición en Psicología. Por otra parte, hay que señalar, que Binet y Simon desarrollaron *procedimientos estandarizados* de medición que permiten controlar los factores perturbadores que pueden influir en la ejecución de los sujetos en los tests. Finalmente estos autores hacen la primera calibración de los *items* reuniendo sistemáticamente datos empíricos acerca de su dificultad en grupos de distintas edades, de manera que trabajando sobre los datos obtenidos en poblaciones bien definidas, establecen las propiedades de los items y definen la edad mental de los sujetos (Barbero, 1996).

El trabajo de Binet y Simon tenía una clara orientación práctica; necesitaba de un marco teórico que diera *fundamentos científicos* a las puntuaciones de los tests. Spearman (1904a, 1904b, 1907, 1910, 1913) fue quien desarrolló el marco teórico de la *Teoría de los Tests*, proporcionando un modelo conceptual (definido

como lineal y aditivo), y una teoría psicométrica, que permitió, a los constructores de tests, trabajar dentro de los parámetros científicos de la época.

E. L. Thorndike con su libro *An introduction to theory of mental and social measurements*, y otras publicaciones posteriores puede ser considerado como: el organizador del primer cuerpo de teorías psicométricas, el que refuerza los estudios empíricos sobre la teoría de la medición, el que amplía los sistemas de interpretación de las puntuaciones de los tests, y el que más contribuye con su polémica con Spearman al esclarecimiento del constructo inteligencia.

La Teoría de los Tests iniciada por estos tres autores, ha sido ampliada, mejorada y diversificada a lo largo del presente siglo, de modo que se puede decir, que históricamente se han observado cuatro sistematizaciones:

- La Teoría Clásica de los Tests (TCT)
- La Teoría de los Tests Referidos al criterio (TRC o ERC)
- La Teoría de las Respuestas a los Items (TRI)
- La Teoría para la Nueva Generación de Tests (TNGT)

Las tres primeras teorías (TCT, TRC, TRI) podrían ser calificadas como cuantitativas, en cuanto que trabajan sobre la eficiencia de los sujetos, expresada en las puntuaciones acreditadas a las respuestas a los items. La última teoría (TGNT), que está surgiendo en nuestros días, podría ser calificada como cualitativa, en cuanto que, basada en el análisis de los patrones comportamentales de los tipos de respuestas dadas a los items del test, de los contenidos de los items, de las frecuencias de los valores vectoriales asignados a los items y del tipo de errores presentes, busca estudiar las estructuras conceptuales, los procedimientos y las estrategias subyacentes, asi como analizar las deficiencias o malas conceptualizaciones en los aprendizajes.

Si concretizamos esta clasificación de las teorías en la medición del rendimiento escolar, se puede afirmar que: las teorías cuantitativas, buscan conocer cuántos conocimientos tiene el sujeto, en cambio las teorías cualitativas intentan ver más bien, cómo los estudiantes piensan, preforman los conocimientos y los aprenden, es decir, la calidad del aprendizaje (ver Díaz, 1995, 1997).

Teoría Clásica de los Tests

La TCT pone su mayor énfasis en las puntuaciones de los sujetos al test, concebidas de este modo:

$$X_{j.} = \sum_{i=1}^{n} x_{ji}$$

El nucleo fundamental de la TCT lo constituyen tres elementos:

- **A.** La concepción de la puntuación del test ($\mathbf{X}_{j.}$), como la suma de los dos componentes constructuales: la puntuación verdadera (V) y la puntuación derivada de los efectos del error (e), y sus derivaciones matemáticas sobre esperanzas, varianzas y correlaciones:

$$X_j = V_j + e_j$$

$$E(V) = E(X) = \mu_X = \mu_V$$

$$\rho^2_{XV} = \sigma^2_V / \sigma^2_X$$

- **B.** El supuesto de que dos tests paralelos deben tener la misma puntuación verdadera, pero que las diferencias halladas en un sujeto son debidas a los efectos del error. La desviación

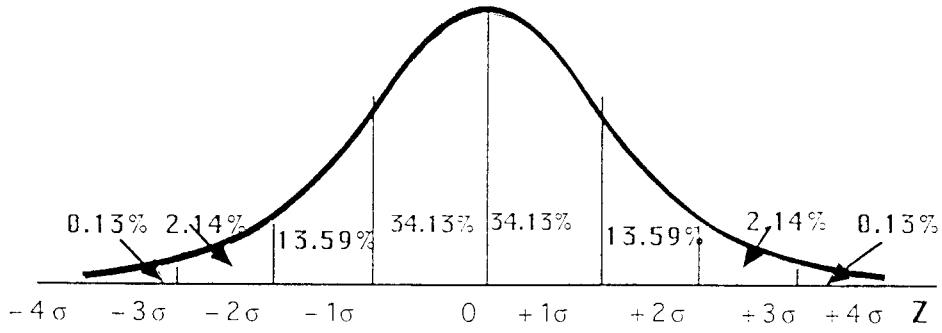
típica de todos los errores se llama el error típico de medida, un concepto esencial en esta teoría:

$$\rho_{XX'} = \rho^2_{XV} = \sigma^2_V / \sigma^2_X$$

$$\sigma_e = \sqrt{\sigma^2_X [1 - \rho_{XX'}]}$$

- **C.** El tercer supuesto se basa en que la distribución de frecuencias de las puntuaciones de los tests, cuando el instrumento está bien construído y se aplica a la población o una muestra representativa y suficientemente grande, sigue el modelo de Gauss.

La curva normal



La Teoría Clásica de los Tests ha dominado durante la primera mitad del presente siglo, y ha sido la base para la construcción de la mayoría de los tests de aptitudes actuales y algunos de rendimiento.

Siguiendo la misma linea sobre el análisis de las puntuaciones de los tests, surge otra teoría, la de la Generalizabilidad de los Tests (TGT), como una forma distinta de estudiar y analizar la fiabilidad y la varianza de las puntuaciones.

La interpretación de las puntuaciones hace referencia a la posición que el sujeto tendría si perteneciera a la población previamente evaluada (Gulliksen, 1950; Lord & Novick 1968)

Teoría de los Tests Referidos al Criterio

Al pretender aplicar los fundamentos de la TCT a la evaluación educativa, los teóricos del área se han visto obligados a crear nuevas técnicas y métodos para evaluar la situación de los sujetos con respecto a algún dominio de aprendizajes bien definidos (Popham, 1978), donde las puntuaciones de lo sujetos hacen referencia a dicho dominio. Este conjunto de técnicas y métodos han sido denominadas globalmente como Tests Referidos al Criterio (TRC), o Evaluación Referida al Criterio (ERC) y se centra en evaluar todos los aspectos del proceso enseñanza-aprendizaje (Hambleton y Rogers, 1991; Martínez Arias, 1995).

Esta modalidad de construir tests se desarrolló en la década de los setenta, y puede ser vista como la introducción del neoconductismo de Skinner en Educación. Estos tests hacen su análisis e interpretación teniendo en cuenta la cantidad de contenidos que el sujeto logra completar. Los tests construídos bajo este enfoque buscan determinar en qué proporción los estudiantes han alcanzado los objetivos operacionales educativos previamente definidos.

Los TRC se fundamentan en los siguientes supuestos:

1° El dominio del test debe ser, al menos teóricamente, definible en sus limites; los items que componen el test deben ser seleccionados aleatoriamente de este dominio.

2° Los items deben considerarse intercambiables, lo que implica que un conjunto combinado de items puede ser sustituido por otro conjunto sin perjuicio de la estimación del dominio del atributo en un sujeto, ya que estas selecciones deben ser equivalentes entre sí y semejantes en sus resultados al supuesto que se aplicaren al sujeto todos los items del dominio.

3° La puntuación del dominio, así estimada, tiene sentido dentro de la concepción muestral de los items, y no como representación del porcentaje de respuestas correctas, que da información solamente acerca del contenido de los items correctamente contestados (De la Orden, 1989).

Los TRC han sido diseñados básicamente para ser utilizados en educación, no como una especie de técnica analítica, sino más bien como instrumentos *de la clasificación sujetos y de medida del progreso educativo*. Para alcanzar estos objetivos se hace necesario determinar estándares y/o puntos de corte.

Esta modalidad de construir tests, aunque se asemeja a la de la clásica en la forma de calificar las respuestas (**eficiencia**), se diferencia profundamente de la misma, de modo que se puede decir que presenta técnicas y componentes propios, por lo que es merecedora de ser considerada como una teoría importante en la medición psicológica y educativa. Las diferencias entre ambas teorías se evidencian sobre todo en estos aspectos:

- **A.** Respecto a la finalidad de los tests, ya que los TRC buscan determinar el estado actual de rendimiento o dominio de contenidos de cada sujeto, para clasificarlo (vgr: "apto", "no apto"), mientras que los TRN buscan situarlo en un continuo que representa el atributo medido.

- **B.** Respecto a la *construcción del test y especificación de los contenidos*, ya que el procedimiento varía sobre todo: en los planteamientos iniciales, en la revisión de los objetivos, estudio de las cualidades de los tests, etc.

- **C.** Respecto a la *selección de los items*, ya que estos se eligen en función de los objetivos buscados con el uso de los tests, y no en función de los valores estadísticos de los items obtenidos con los resultados del grupo sobre el cual se hace el estudio piloto, como en la TRN.

- **D.** Respecto al *significado de las puntuaciones*, pues las mismas son consideradas como un estimador muestral de rendimiento en el dominio, y no como un estimador de las puntuaciones verdaderas de los sujetos, y

- **E.** Respecto a la *interpretación de las puntuaciones*, ya que éstas tienen un significado en términos absolutos, en cambio en la TRN tienen un significado relativo.

Al contrario de lo que sucede en la TCT las puntuaciones de los TRC no necesitan ser transformadas puesto que tienen sentido en sí mismas, ya que éstas se expresan generalmente en porcentajes de dominio que tiene el segundo. Esta información debe ser completada indicando también las areas del dominio que el sujeto domina y no domina.

Estos tests resultan útiles para evaluar el aprendizaje escolar, ya que proporcionan criterios sólidos para decidir si un estudiante supera o no suficientemente una tarea u objetivo. Su utilidad crece aun más, cuando se trata de evaluar una área de instrucción concreta o el logro de objetivos básicos de un programa, y sobre todo para la evaluación continua de los logros que los sujetos van o no alcanzando a medida que avanzan en el proceso enseñanza-aprendizaje.

Aunque los TRC se han aplicado básicamente a los tests de rendimiento académico, también han sido aplicados en otros campos como: en las Fuerzas Armadas para determinar el grado de competencia, en la industria para señalar las destrezas de los sujetos para colocarlos en un puesto de trabajo concreto, y en el área clínica para medir los cambios relativos a la aplicación de tratamiento, y a lo mejor también en las denominadas evaluaciones conductuales.

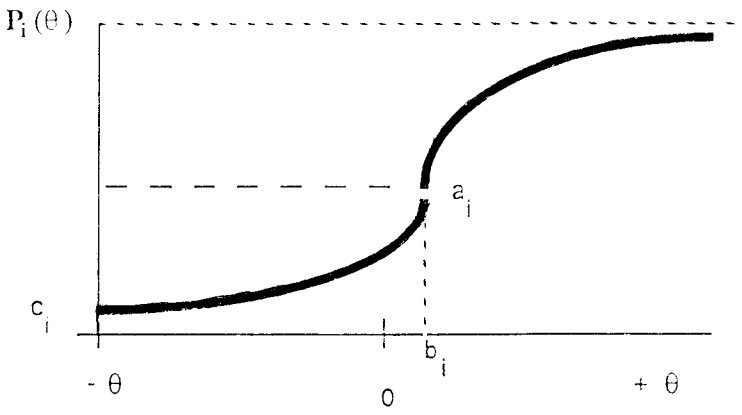
Esta teoría podría ser aplicada, quizás, a la medición de aptitudes. Si esto se lograra se tendría un gran avance en Psicometría, pero la gran dificultad está en la posibilidad de establecer el universo de estos contenidos, cuando no se conocen bien los constructos y las estructuras o funciones subyacentes en las mismas, asunto que intenta solucionar la Psicología cognitiva.

Teoría de las Respuestas a los items

La TRI se basa en la estimación de la probabilidad de acertar a los items condicionada a la aptitud del sujeto.

$$P(\overline{X} \mid \theta)$$

Esta condición permite a esta teoría trabajar sobre una matemática mucho más sólida, como es la teoría de las probabilidades. Esta función de probabilidad suele visualizarse sobre la llamada curva característica del item (CCI):



En la que se pueden apreciar tres parámetros:

- El $\mathbf{b}_i$ llamado parámetro de dificultad cuyos valores pueden estar entre:

$$-\infty \leq b_i \leq \infty$$

- El parámetro $\mathbf{a}_i$ llamado parametro de discriminación, sus valores pueden estar entre:

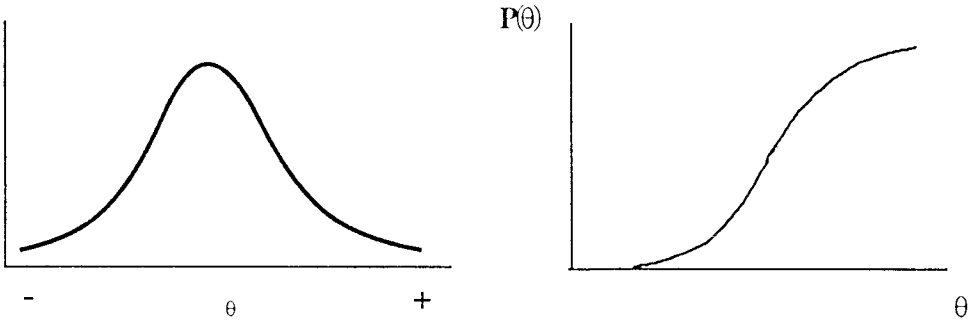
$$0 \leq a_i \leq \infty$$

- El parámetro $\mathbf{c}_i$, llamado parámetro de conjetura o seudo-azar, sus valores están entre:

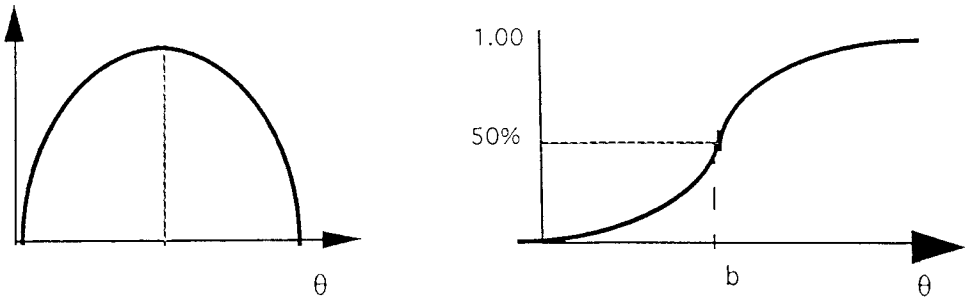
$$0 \leq c_i \leq 1.0$$

La TRI acepta cualquier tipo de modelo de CCI que se ajuste a los datos. No obstante los más comunes son los de la ojiva normal y los logísticos:

Modelo de la ojiva normal



Modelo de logístico



Cada uno de estos modelos a su vez se subdivide en otros. cuyas funciones y cuyas expresiones son:

Modelos de la ojiva normal

1P : $P_1(\theta) = \int_{-\infty}^{(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{(-z^2/2)dz}$

2P: $P_1(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{(-z^2/2)dz}$;

3P: $P_1(\theta) = c_i + (1 - c_i) \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{(-z^2/2)dz}$

Modelos logísticos

$$1P.\quad P_i(\theta) = \frac{1}{1+e^{-D(\theta-b_i)}};$$

$$2P.\quad P_i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_i)}};$$

$$3P.\quad P_i(\theta) = c_i + (1-c_i)\frac{1}{1+e^{-Da_i(\theta-b_i)}}$$

Para poder operar con estos modelos es nesario que se cumplan ciertas condiciones como:

- la **independencia local** de los items, de los sujetos y de las aptitudes,
- la **unidimensionalidad** del rasgo, y
- la **falta de presión temporal** en la ejecución del test.

Si estas condiciones no se dan, hay que trabajar con otros modelos, como pueden verse en el libro de Van der Linden y Hambleton, 1997.

El punto neurálgico de la TRI es la estimación de los parámetros. Una buena estimación de los mismos solo es posible hacerla a través de un ordenador que utiliza algoritmos de optimización como el EM y el Newton-Gauss que resuelven las ecuaciones que optimizan los resultados.

Los programas más utilizados son:

BICAL	Wright *et al.* (1970)
BILOG	Mislevy y Bock (1990)
MULTILOG	Mislevy y Bock (1988)
LOGIST	Wingersky, Barton y Lord (1984)
DATAGEN	Hambleton y Rovinelli (1983)

Nosotros utilizamos preferentemente el BILOG 3, que cubre tres fases:

- en la primera calcula los valores de los parámetros de los items según la TCT.
- en la segunda se hace una calibración de los items, que permite ofrecer los valores paramétricos estimados y ajuste del modelo.
- en la tercera se presentan las funciones de información de los items, las puntuaciones de los sujetos, las características del test, así como la distribución de la Información en la Población.

No cabe la menor duda que la TRI resuelve mucho mejor que la TCT problemas psicométricos, como:

- en la construcción de tests, ya que si se conocen sus parámetros, la selección y sustitución de items es mucho más fácil, así como la determinación del error y la precisión del test (Theunissen, 1985).
- en la elaboración de los tests adaptativos, ya que se pueden escoger los items más próximos a la aptitud del sujeto (Weis, 2984).
- en la evaluación educacional en gran escala, ya que puede medir mejor la eficiencia al nivel poblacional, asi como evaluar programas o estrategias al poder disponer de instrumentos capaces de hacer estimaciones, invariadas y estables (Bock, Mislevy y Wodson, 1982; Messick, Beaton y Lord, 1983).
- las técnicas de la TRI parecen tener más potencial que el demostrado en la actualidad, ya que se han utilizado generalmente a resultados de tests, concebidos, construídos y administrados dentro del marco teórico clásico donde los items no contestados no se tienen en cuenta (Mislevy, 1991).
- el TRI al trabajar sobre los patrones de respuestas, puede efectuar análisis de respuestas inesperadas o raras que

pueden revelar concepciones defectuosas o imprecisas en el aprendizaje (Tatsoucka, 1983).

- finalmente, entre otras cosas, el TRI permite avanzar con más seguridad en el estudio de los automatismos intelectuales que el sujeto utiliza para adquirir conocimientos, de las estrategias mentales que utilizan los sujetos (destrezas metacognitivas), de la construcción de esquemas que relacionan hechos y destrezas, etc.

Teoría para una Nueva generación de Tests

En los inicios de los años noventa aparece un grupo de autores que intentan integrar las nuevas concepciones e investigaciones de la Psicología cognitiva con las teorías y técnicas de medición vigentes (la TRI) para recrear una nueva teoría de los tests y desarrollar nuevos diseños de los tests, ya que las teorías y diseños existentes resultan insuficientes para explicar la complejidad del comportamiento cognitivo y del proceso de aprendizaje. Este intento integrador ha sido denominado: Teoría de los Tests para una Nueva Generación de Tests (TNGT, Frederiksen, Mislevy, y Bejar, 1993)

Este movimiento más que una disciplina ha de ser considerado, como una especie de confederación de científicos (psicólogos cognitivos, psicólogos educativos y educadores y psicólogos metodólogos), que están estudiando nuevos aspectos del comportamiento cognitivo o del aprendizaje.

Dos son, pues, las lineas de trabajo en que se ha desarrollado esta teoría: el análisis de las aptitudes cognitivas y el rendimiento académico, lineas que se ajustan a los dos tipos de profesionales que están participando en su creación.

Las aptitudes cognitivas concebidas hasta ahora como unidad inobservable (rasgo latente), comienzan a ser visualizadas como una especie de configuración o constelación individual, en la que

se integran e intercorrelacionan: conocimientos, niveles de comprensión, estrategias de trabajo, destrezas, creencias y actitudes. Esta configuración, en realidad, no es otra cosa, que el resultado mutante de la integración de las experiencias y aprendizajes, que cada sujeto va haciendo de acuerdo a su idiosincracia biogenética y ambiental (Lohman y Ippel, 1993). Otro aspecto importante de la actividad mental, que la TNG intenta solucionar es la identificación de los procesos mentales que subyacen en los comportamientos mentales.

La existencia predominante de psicólogos educativos y educadores ha hecho que muchas de sus investigaciones de la TNGT se hayan centrado en el estudio del aprendizaje académico percibido por estos autores como un aprendizaje activo, donde el sujeto construye conscientemente sus propias comprensiones, conocimientos y relaciones (Di Vista,1989). El estudio de este tipo de aprendizaje implica: describir su naturaleza, lo que equivaldría a señalar las reglas que utilizan los estudiantes en la formación de conceptos y en la representación de los materiales que intervienen en el proceso de aprendizaje.

De modo que, según esta teoría el objetivo de los tests educacionales será, más que conocer cuántos conocimientos posee el sujeto, comprobar su estado de competencia , esto es, conocer el estado o nivel de desarrollo en que se encuentra la constelación de conceptos y destrezas de los sujetos con respecto a una disciplina. El punto neurálgico de estas investigaciones está en apreciar y comprobar las diferencias existentes entre los estudiantes "iniciados o noveles y los avanzados o expertos" sobre los procedimientos y heurísticas conscientes del sujeto, sobre las estructuras de conocimientos, sobre las técnicas y estrategias utilizadas para resolver los problemas, así como sobre la capacidad de focalizar sus esfuerzos. De modo que la diferencia entre ellos reside esencialmente en la formación y uso de modelos apropiados o no a las diversas situaciones de medición optadas: estudio de la físi-

ca (Chi, Feltovich y Glasser, 1981), el aprendizaje del ajedrez (Chase y Simon, 1975); el estudio de la radiología (Lesgold, Feltrovich, Glasser y Wang, 1981), el estudio de las ciencias sociales (Woss, Greene, Post y Penner, 1983), etc.

La TNGT se fundamenta sobre estos supuestos:

1°. Al igual que en las otras teorías, ésta supone que detrás de las respuestas de los sujetos a los items existe una aptitud, que los sujetos la poseen de distinta forma y que al fin y al cabo es la responsable de los diferentes tipos de respuestas. A pesar de esta conceptualización común a ambos grupos de teorías (cuantitativas y cualitativas), existe una diferencia entre ellas, porque las primeras conciben estas aptitudes como una unidad funcional e impenetrable del ser humano (rasgo latente), en cambio la nueva teoría intenta desentrañar su naturaleza y descomponer sus elementos, mediante definiciones operacionales más accesibles (Haertel y Wiley).

2°. Los instrumentos de medida o test no son conceptualizados como una muestra representativa de los items homegeneos que forman el dominio de la aptitud, sino un conjunto de actividades no necesariamente homogeneas, desarrolladadas para medir distintos aspectos de la aptitud claramente definida en su aspecto cualitativo y en su aspecto cuantitativo.

3°. Esta nueva teoría se fija en todos los aspectos de las respuestas de los sujetos, y no solo en su eficiencia como lo han hecho las teorías vistas hasta ahora. De modo que en las mismas distingue dos aspectos: uno cualitativo (el contenido de los items o componentes, los procesos de información y las estrategias que el sujeto utiliza para contestarlas), y otro cuantitativo que corresponde a las valoraciones de estos elementos, en particular, al estudio de los patrones de respuestas observados tanto en algunos subgrupos de items (testslets) como en el test total. De todos modos hay que resaltar que los teóricos de esta teoría están más interesados en la búsqueda

de la forma como los sujetos alcanzan las respuestas, que en la cuantificación de las mismas.

Esto implica considerar la variabilidad intraindividual, objetivo final del uso de los tests, como la resultante de las diferencias en el procesamiento de las distintas tareas que definen la aptitud, como proponía Guttman (1971).

Para llevar a cabo el estudio de esta variabilidad la TNGT propone cumplimentar dos elementos:

1°. Hallar un modelo de procesamiento, después de hacer un análisis conceptual controlado de distintas ejecuciones de tareas, que permita señalar los componentes esenciales y los pasos (*steps*) del proceso de actuacion del sujeto en los items y en el test, y

2°. Elaborar un modelo de medición, que permita asignar valores adecuados a las ejecuciones de los diversos componentes.

Lo que nos lleva a recalcar que en la aplicacion de esta teoría a cualquier atributo deben estar presentes dos partes: una sustantiva que cubriría el primer paso y otra técnica que abarcaría el segundo (Bejar).

De modo que el procedimiento general propuesto para trabajar con este tipo de teoría, cubriría las siguientes etapas:

- efectuar, antes que nada, un análisis minucioso de las tareas o subtareas (*task analysis*) que hay en cada test, siguiendo las normas de la psicología cognitiva, para obtener estructuras cognitivas capaces de explicar los comportamientos intelectivos o de aprendizaje a estudiar,
- utilizar los diseños de tests adecuados que permitan diferenciar mejor las partes del procesamiento intelectivo,
- finalmente fundamentar la medición en un nueva teoría de los tests, pues las existentes (TCT y TRI) no resultan ade-

cuadas a los fines propuestos, ya que se fundamentan en modelos muy simplificados de las aptitudes limitándose a realizar algunas aplicaciones estadísticas a las puntuaciones globales de los tests o de los items (Lohaman e Ippel, 1993).

Esta concepción compleja del objeto de la medición (ya sean las aptitudes o el redimiento académico) lleva a proponer modelos de medición más complejos, cuyos parámetros sean capaces de medir y explicar los patrones encontrados y los procesos mentales inferidos.

Modelos

Así frente al modelo clásico de diseño dicotómico de la TRI antes indicado, Thissen (1993) y Bennet (1993) proponen buscar diseños más complejos que permitan medir los distintos niveles de atributo. En tal sentido proponen cuatro tipos de modelos:

- los de respuestas nominales,
- los de respuestas graduadas,
- los de respuestas de selección múltiple,
- los de crédito parcial sobre tareas complejas,

modelos que no cumplen plenamente los supuestos de la TRI.

Los **modelos de respuestas nominales**, propuestos inicialmente por Bock (1972) buscan construir un trazado lineal (una especie de CCI) para cada una de las categorías del item (o pseudoitem) y hallar la probabilidad de obtener un patrón de respuestas (x), mediante las siguientes formulas:

$$T_k(\theta) = \frac{1}{1+e^{-a_k(\theta-b_k)}}; \quad T_x(\theta) = \sum_{k=1}^{m} \frac{1}{1+e^{-a_k(\theta-b_k)}};$$

$$P(x) = \int_{-\infty}^{+\infty} \prod_{i=1}^{nitems} T_{xi}(\theta)\, \phi(\theta)\, d\theta$$

Donde:

- a y b son los parámetros de discriminación y de dificultad para k = 1, 2, 3, ..., m categorías;
- $\Phi(\theta)$ la distribución de la aptitud latente en la población, que se asume que es normal.

Dentro de este modelo Thissen (1993) cita varios estudios, hechos según las perspecivas de la TNGT:

- el de Bergan y Stone (1985) que pretende medir el dominio de los numeros 3 y 4 en los niños de preescolar, distinguiendo dos componentes intelectivos: las capacidades de identificarlos o asociarlos a grupos de objetos; a su vez distingue dentro de cada componente cuatro categorías (ninguno, solo el 3, solo el 4, ambos) (pp.80-6).

- el de Klassen y O'Connor (1987) que busca efectuar un estudio perspectivo de los predictores de la violencia en adultos varones que ingresan los centros de salud mental, tomando como variables las edades (más de 34, entre 25 y 34, entre 18 y 24 y menos de 18), en cada una de ellas establece cuatro categorías sobre las veces que han sido ingresados (ninguna, de 1 a 3, de 4 a 9, 10 o más veces) (pp. 86-89), y

- el de Irving (1987) sobre un cuestionario de 36 items sobre Bulimia denominado BULIT, que pretende medir los riesgos de adquirir esta enfermedad, en base a la presencia de diversos desórdenes en la comida, en los que dintinguen a su vez cinco alternativas.

Los **modelos de respuestas graduadas** (tipo likert), propuestos inicialmente por Samejima (1969), están organizados de modo que el orden de las respuestas está condicionado a la capacidad o nivel de aptitud del sujeto. Estos modelos presentan un trazado lineal para respuesta ordenada (x = k) del item, así como los patrones de respuestas, utilizando fórmulas muy semejantes a la anterior

$$P(x = k) = \frac{1}{1 + e^{[-a(\theta - b_{k-1})]}} - \frac{1}{1 + e^{[-a(\theta - b_k)]}}$$
$$= P^*(k-1) - P^*(k)$$

Donde:

- a es la discriminación del item y b la dificultad para k = 1, 2, 3, ... m de las respuestas graduadas,
- P*(K) es la probabilidad de que una respuesta sea k o mayor para cada valor de θ.
- b_{k-1} es el valor de localización en el eje de abcisas θ de la probabilidad de 50% de que el sujeto obtenga el valor k o mas alto.

Varios de los items del cuestinario del BULIT presentan este diseño de item.

Los **modelos de respuestas de selección múltiple**, basados en algunas modificaciones de modelos anteriores (Samejima, 1969; Bock, 1972, es propuesto por Thissen y Steinberg (1984, 1986) para analizar los items de selección múltiple en todas las respuestas presentadas (clave y distractores), presentando trazados lineales para cada alternativa de respuesta (k = 2, 3, 4, ..., m+10), utilizando la siguiente fórmula:

$$P(x = k) = \frac{h^* e^{(a_k\theta + c_k)} + h\, d_k\, e^{(a_i\theta + c_i)}}{\sum_{i=1}^{m+1} e^{(a_i\theta + c_i)}}$$

Donde:

- a y c no están relacionados con la localización, sino con el contraste entre los parámetros estimados hecho por cual-

quiera de estos tres valores matriciales: desviaciones, polinomiales o triangulares.

- el parámetro d_k representa la proporcion de la alternativa 1, que se aplica a los que no contestan al item (*don't know*), con esta fórmula:

$$d_k = \frac{e^{\left[d_k^*\right]}}{\sum e\left[d_k^*\right]}$$

Estos modelos y algunos otros más derivados de los anteriores pueden ser calculados prácticamente mediante el uso del programa MULTILOG de Thissen.

Los **modelos de credito parcial o de respuestas construidas**, no piden al sujeto que identifique la respuesta correcta, sino que formule los pasos que le llevan a dar la respuesta. Este tipo de respuestas complejas busca identificar la secuencia de las activides mentales del sujeto, para expresar el nivel de desarrollo o los defectos de concepción o ejecución de los sujetos, basados en un esquema secuencial teórico. La dificultad mayor de este tipo de items es la calificación de estas respuestas complejas. Esta calificación puede ser hecha por expertos, método que se intentó en el *College Borad's Advanced Placement Program*, y que resultó sumamente caro (ya que los jueces debían ser entrenados, trasladados, alojados y pagados) y largo (ya que debían calificar miles de exámenes y preguntas), por lo se está intentando hacer programas que ayuden a esta tarea, como: el APCS (*Advanced Placement Computer Science*, de Bennet, Sack y Soloway, 1991), el GRE (*Graduate Record Examination*, de Bennet, Sebrechts, y Yamamoto 1991). Dentro de esta línea se han desarrollado otros programas específicos, unos para detectar los errores de concepción en distintas áreas tales como: el PROUST y MicroPROUST desarrollados para detectar los errores conceptuales a el aprendizaje

de la programación en Pascal (Johnson, 1986, Johnson y Soloway, 1985); Braun, Bennett, Frye, y Soloway, 1990), el GIDE creado para detectar errores en la solución de problemas de algebra (Sebrechts, LaClaire, Schooler y Soloway, 1986; Bennett, Sebrechts, y Rocks, 1991).

Para detectar el nivel de eficiencia o desarrollo de los sujetos, se han desarrollado, dentro de esta misma línea, otros modelos tales como el HOST (*Hierarchically Ordered Skills Tests*, de Rock y Pollack, 1987) y el HYBRID (Yamamoto, 1987) y el Modelo Master de crédito parcial (Master y Mislevy, 1993).

De manera que los tests ya no deben ser concebidos como una muestra representativa del dominio, sino como un conjunto estructurado de items, seleccionados de acuerdo al modelo de procesamiento elegido (Ippel, 1986).

Los patrones de respuestas a los items

Frente al clásico análisis de items, la TNGT pone su máximo esfuerzo en el estudio de los patrones de respuestas a los items previamente calibradas en su dificultad por cualquiera de los modelos propuestos dentro de la TRI según la naturaleza de los items utilizados: dicotómicos (Rasch, Birmaum, Lord, Bock, Wright), linealmente politomos (Haladyana y Sympson, 1988; Mellenbergh, 1995,1996) y los politomos (Bock, 1972; Samejima, 1979; Sympson, 1983, 1986; Thissen y Steinberg, 1986). De manera que cada sujeto no es caracterizado por la puntuación que recibe por número de items contestados, sino por el patrón de respuestas, en el que se pretende ver las tendencias a actuar de los sujetos de acuerdo a su estado de competencia, así como los tipos de errores que comete. Variable que puede ser considerada como categórica o continua.

Bajo esta perspectiva se puede afirmar que detrás de los valores paramétricos de los items se perciben varios tipos factores de influencia:

- los debidos a las características psicológicas del sujeto, tales como: la capacidad de atención, la velocidad de ejecución y el sentido de precisión, el nivel de stress emocional, la transferencia que otorga el sujeto a la tarea de responder al test, a la fatiga mental, etc.;
- los debidos a las condiciones del item, como su posición en la secuencia, el formato, su relación con la aptitud, su capacidad de discriminación, etc.;
- los debidos a la etapa de desarrollo en que el sujeto se encuentra, tales como: el de adquisicion donde se hacen los primeros contactos conceptuales, el de procesamiento que incluye la elaboracion e interconexión de conceptos o el de automatización, que implica reestructurar los conceptos nuevos y viejos en esquemas mentales que agilizan las ejecuciones;
- los debidos a las estrategias utilizadas por el sujeto para alcanzar las respuestas;
- y finalmente, los debidos a la capacidad del sujeto de adaptación a las tareas (es decir el aprendizaje logrado durante la ejecución del test) y a la flexibilidad mental para buscar nuevas soluciones.

En el análisis de los patrones de respuestas se han propuesto tres modelos de asociación de items: el modelo de las placas tectónicas (Wilson, 1985,1989), el modelo de las clases latentes (Haertel, 1984), el modelo de los componentes (Embretson 1983,1985b; Samejima, 1983).

Reflexiones finales sobre la TNGT

Sintetizando se puede decir que esta teoría focaliza su actividad:

- en el analisis de las diversas subtareas que realiza el sujeto.

- en la busqueda de las estrategias de solución que utiliza el sujeto.
- y en el estudio de los patrones de las respuestas, en los que toma en cuenta tanto los aciertos (1) como los fracasos (0), así como las causas de estos fracasos: las malas conceptualizaciones, los fallos en las estrategias, los defectos en la ejecución, etc.

La nueva teoría, en definitiva, se basa en el análisis de los conceptos claves y sus interacciones, en la velocidad de respuestas, y en las estructuras de conocimiento, no descritas por expertos, sino halladas en los contrastes efectuados entre grupos competentes y no competentes.

La integración todos estos elementos en la medición psicológica resulta una tarea ardua y excitante. Para conseguir estos elementos a través de los tests es necesario utilizar mucha imaginación y creativad para desarrollar nuevos tipos de items y nuevos modelos estadísticos, que rompan las amarras que han tenido estática a la Teoría de los tests. La revolución que hay que realizar no es pequeña, ya que hay que recrear la teoría de los tests (Cole, 1993).

REFERENCIAS BIBLIOGRÁFICAS

Barbero, Isabel (1996). "Bancos de Items". En J. Muñiz (Ed.) *Psicometría*. Madrid. Universitas Bennet, R. E. y Ward, W.C. (Eds) (1993): *Construction versus choice in cognitive measurements: Issues in constructed response performance testing and porfolio assesment*. Hillsdale, London. Lawrence Erbaum.

Bergan, J. R. y Stone, C. A. (1985). "Latent class models for knowledge domains." *Psychological Bulletin*, **98**, 166-184.

Binet, A. & Simon,T. (1905). "Methodes nouvelles pour le diagnostique du niveau intellectuel des anormaux." *Année Psychologique*, **11**, 191-244.

Binet, A. & Simon,T. (1908). "Le developpement de l'intelligence chez les enfants." *Psychologique*, **14**, 1-94

Binet, A. (1911). "Nouvelles Recherches sur la measure du niveau intellectuel chez les enfants d'echole." *Revue Philosophie*, **11**, 191-244.

Bock, R. D. (1972). "Estimating item parameters and latent ability when responses are scored in two o more categories." *Psychometrika*, **37**, 29-51.

Bock, R. D. Mislevy, R. J. y Woodsen, C. E. M. (1982). "The next stage in educational assessment." *Educational Researcher*, **11**, 4-11, 16.

Chase, W. G. y Simon, H. A. (1975). "Perception in Chess." *Cognitive Psychology*, **4**, 55-81.

Chi, M. T. H., Feltovich, P. y Glasser, R. (1981). "Categorization and representation of phisics problems by experts and novices." *Cognitive Sciences*, **5**, 121-152.

De la Orden, A.(1989). "Investigación cuantitativa y medida en educación." *Bordon*, **41**, 217-2

Díaz, J. V. (1995): *Construcción de Tests*. Psicometría I y II. Valencia, Cristóbal Serrano.

Díaz, J. V. (1997): *La Teoría de las respuestas de los items aplicada a la Construcción de Tests de aptitudes*. Valencia. Cristobal Serrano.

Díaz-Aguado, M. J. (1997). "Las relaciones interpersonales en la escuela." En: F. Rivas. *El proceso de ensenanza/aprendizaje en la situación educativa* (Epílogo). Barcelona, Ariel.

Embretson, S.E (1993). "Psychometrics models for learning and cognitive process." In: N. F. Frederiksen, R.J. Mislery and I.I. Bejar (Eds). *Test theory for a new generation of tests*. Hillsdale, N. J. Lawrence Erlbaum.

Embretson, S. E (Ed.) (1985). *Test design: Devolopment in psychology and psychometrics*. N. Y. Academic Press.

Frederiksen, N., Mislevy, R. J. y Bejar, I. I. (Eds) (1993). *Test theory for a new generation of tests*. Hillsdale, N. J. Lawrence Erlbaum.

Gulliksen, H. (1950, 1987). *Theory of Mental Tests*. Hillsdale N. J. Lawrence Erlbaum

Haladyna, T. M. y Sympson, J. B. (1988). "Empirically based polychotomous scoring of multiple-choice test items: A review." In: *New Develoment in Polychotomous Scoring. Symposium conducted at the annual meeting of the American Educational Research Association*, New Orlens.

Hambleton, R. K. y Swaminathan, H. (1985). *Item response theory: Principles and Application*. Boston, Kluwer-Nijhoff Publishing.

Hambleton, R. K y Rogers, H. J. (1991). "Advances in criterion-referenced measurement." En: R. K Hambleton y J. N. Zaal (Eds.). *Advences in educational and psychological testing: Theory and applications*. Boston, Kluver.

Hambleton, R. K., Zaal, J. N. y Pieters, J.M. (1991). *Advances in educational and psychological testing, theory and applicacions*. Boston: Kluwe.

Heartel, E. H. (1984). "An application of latent class models to assessment data." *Applied Psychological Measurement*, **8**, 333-346.

Heartel, E. H. y Wiley, D. E. (1993). "Representations of Ability Structures: Implications for Testing." En: N.F. Frederiksen, R Mislery and I.I. Bejar (Eds). *Test theory for a new generation of tests*. Hillsdale, N. J. Lawrence Erlbaum.

Ippel, M. J. (1986). *Component-testing: A theory of cognitive aptitude measurements*. Amsterdam: Free University Press.

Irving, L. M. (1987). *Mirror images: Effects of the standard of beauty on women's self and body esteem*. Unpublished master's thesis, University of Kansas.

Klassen, D. y O'Connor, W. A (1987). "Predicting violence in mental patients: Cross-validation of an actuarial scale." Paper presented at the annual meeting of American Public Health Association, New Orleans.

Lesgold, A. M., Feltovich, P. J., Glaser, R., y Wang, Y. (1981). "The acquisition of perceptual diagnostics skill in radiology." (Tech. Rep. N° PDS-1).

Lohman, D. F. y Ippel, M. J. (1993). "Cognitive Diagnosis: From Statistically Based Assessment Toward Theory-Based Assessment."

In: N. Frederiksen, R. J. Mislevy y I. I. Bejar (Eds). *Test Theory for a New Generation of Tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M. and Novick M. R. (1968). *Statistical of theories of mental test scores*. Reading Mass. Addison-Wesley.N.Y.

Martínez Arias, R. (1991). "Inteligencia y procesos superiores." En: R. Martínez Arias y M Yela (Eds). *Pensamiento e inteligencia*. Madrid: Alambra.

Martínez Arias, R. (1995). *Psicometría: Teoría de los tests psicológicos y educativos*. Madrid: Síntesis Psicología.

Mayer, R. (1992). "Cognition and instruction. Their Historic Meeting Whithin Educational Psychology." *Journal of Educational Pschology*, **84**, 4, 405-412.

Mellenbergh, G. J. (1995). "Conceptual noters on models for discrete polytomuous item responses." *Applied Psychological Measurement*, **19**, 91-100.

Mellenbergh, G. J. (1996). "Modelos para items politómicos de respuesta discreta." En: J. Muñiz, *Psicometría* (pp. 786-810). Madrid: Universitas.

Messick, S, Beaton, A. E. y Lord, F. M. (1983). *National Assessment of Educational Progress reconsidered: A new design for a new era*. (NAEP Rep-83-1). Princeton, NJ: National Assessment of Educational Progress.

Mislevy, R. J. (1991). "Randomization-based inferences about latent variables form complex samples". *Psychometrika*, **56**, 177-196.

Múñiz, J. (1990). *Teoría de respuesta a los items: un enfoque en la evaluación psicológica y educativa*. Madrid. Pirámide.

Múñiz, J. (1993). *Teoría clásica de los tests*. Madrid. Pirámide.

Múñiz, J. (1996). *Psicometría*. Madrid. Universitas.

Pellegrino, J. W. (1988). "Mental models and mental tests." En: H. Wainer y H. Braum (Eds.). *Tests validity*. Hillsdale, NJ: Lawrence Erlbaum.

Popham,W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Printice Hall.

Rivas, F. (1997). *El proceso de enseñanza/aprendizaje en la situación educativa*. Barcelona, Ariel.

Ronning, R. R., Glover, J.A., Conoley, J. C. y Witt, J. C. (Eds.) (1987). *The influence of cognitive psychology on testing*. Hillsdale, NJ: Lawrence Erlbaum.

Samejima, F. (1969). "Estimation of latent ability using a response pattern of graded scores." *Psychometrika Manographs*, **34**, (4, Pt. 2, Whole N° 17).

Snow, R. E. (1989). "Implications of cognotive psychology for educational measurement". En R. L. Linn (Ed): *Educational measurement* (3rd Ed., pp. 263-331). Hillsdale, NJ. Lawrence Erlbaum Associates .

Spearman, C. (1904a). "The proof and measurement of association between two things." *American Journal of Psychology*, **15**, 72-101.

Spearman, C. (1904b). "General Intelligence" objectively determined and measured. *American Journal Psychology*, **15**, 201-292.

Spearman, C. (1907). "Demostration of formulate for true measurements of correlations". *American Journal of Psychology*, **18**, 161-69.

Spearman, C. (1910). "Correlations calculated from faulty data". *British Journal of Psychology*, **3** , 271-295.

Spearman, C. (1913). "Correlations of sum of differences". *British Journal of Psychology*, **5**, 417-426.

Stenberg, R. J. (1977). *Inteligence, information processing and analogical reasoning: The componential analysis of human abilities*. Hillsdale, N.J. Lawrence Erlbaum.

Stenberg, R. J. (1988). "Geneces: A rationale for the construct validation of theories and tests of intelligence." En: H. Wainer y H. I Brawn (Eds). *Tests validity*. Hillsdale, N.J. Lawrence Erlbaum.

Stenberg, R. J. (1991). "Cognitive theory and psychometrics." En: R. K. Hammbleton y Y. N. Zaol (Eds.). *Advances in educational and psychological testing. Theory and applicacions*. Boston: Kluwe.

Sympson, J. B. (1986). "Extracting information from wrong answer in computarized adaptative testing." Paper presented at the annual meeting of the American Psychological Association (Washington, D.C.)

Tatsuoka, K. K. (1983). "Rule space: An approach for dealing with misconceptions based on item response theory." *Journal of Educational Measurement*, **20**, 345-354.

Theunissen, T. J. M. (1985). "Binnary programing and test design." *Psychometrika*, **50**, 411-420.

Thissen, D. (1993). "Repeating the rules that no longer apply to psychological measurement (pp. 73-79)." In: N. Frederiksen, R. J. Mislevy y I. I. Bejar (Eds). *Test Theory for a New Generation of Tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D y Steinberg, L. (1984). "A response model for multiple-choice items." *Psychometrika*, **49**, 501-519.

Wilson, M. R. (1985). *Measurement stage of growth: A Psychometric model of hierarchical development* (Ocasional Paper N° 19) Hawthorne, Australia Council for Educational Research.

Wilson, M. R. (1989). "Saltus: A psychometric model of discontinuity in cognitive development". *Psychological Bulletin*, **105**, 276-289.

Voss, J. F., Greene, T. R., Post, T. A. y Penner, B. C. (1983). "Problem-solving skill in the social sciences." In: G. H. Bower (Ed.) *The*