Master en Estudios Avanzados en Celebro y Conducta.

Diseño y medición de programas de intervención neuropsicológica: aspectos fundamentales.

Sevilla, Enero de 2008

Salvador Chacón Moscoso
José Antonio Pérez-Gil
Universidad de Sevilla

CAPITULO 2

La Teoría Clásica de los tests.

2.1. Introducción

2.2. Principales Teorías de los Tests

2.3. La Teoría Clásica de los Tests (TCT)
- 2.3.1. Introducción a la TCT
- 2.3.2. Supuestos fundamentales de la TCT
- 2.3.3. Concepto de tests paralelos: el modelo de formas paralelas y sus variantes
- 2.3.4. Estimación de la puntuación verdadera
  - 2.3.4.1. Estimación mediante el teorema de Chevychev.
  - 2.3.4.2. Estimación basada en la distribución normal de los errores
  - 2.3.4.3. Estimación según el modelo de regresión
- 2.3.5. Estimación del error de medida
- 2.3.6. Ventajas y limitaciones de la TCT
- 2.3.7. Variantes del modelo clásico lineal

2.1. Introducción.

Como ya se ha apuntado en el capitulo anterior, la teoría de los tests surge ligada al estudio de las diferencias individuales; éstas, como señalan Muñiz (1992) y Navas (1997), han sido reconocidas y apreciadas desde antiguo, así los primeros tests o pruebas de aptitud, utilizadas por el gobierno Chino, pueden cifrase aproximadamente en el 2000 antes de Cristo. No obstante, los orígenes que darán lugar a los actuales tests hay que ubicarlos en los primeras intentos de medición de diferencias individuales realizadas por Sir Francis Galton (1822-1911) junto con Alfred Binet (1857-1911) y James Mckeen Cattell (1860-1944).

Galton, primo segundo de Darwin e influenciado por éste, se interesó en la problemática de la herencia y aplicó las ideas de la transmisión de las características físicas al plano psíquico, sobre todo al de la inteligencia. En 1869 publicó su obra Hereditary Genius donde intenta justificar la acumulación de talentos en algunas familias. Para obtener datos sobre diferencias de talentos desarrolló una serie de pruebas, y, en 1882 fundó en Kensington su famoso Laboratorio Antropométrico, equipado para realizar una variedad de tests senso-motores simples; creía que podía medir el intelecto con sus sencillos tests de discriminación sensorial. Fue el primero que aplicó los métodos estadísticos para analizar los datos y no encontrando adecuada la curva normal y sus aplicaciones más simples, inventó un conjunto de herramientas estadísticas adicionales, entre ellas, (con la ayuda de Karl Pearson), el método de correlación, el uso de medidas estándares, la mediana, y métodos de escalamiento psicológicos como el orden-de-méritos y el rating-scale. (Amelang y Bartussek, 1986).

Catell se dedicó también al estudio de la inteligencia; y, como Galton, pensaba que las funciones intelectuales determinan el rendimiento de los órganos sensoriales y que sólo los procesos muy específicos pueden establecerse con precisión, y no los procesos más complejos y "superiores". Fue el primero en utilizar el término Test Mental en su artículo Mental tests and measurement publicado en la revista Mind en 1890. Los tests de Cattell, al igual que los de Galton, centraban su interés fundamentalmente en las diferencias individuales, limitando sus mediciones a un conjunto de índices antropométricos y senso-motores.

La mayor parte de estos primeros trabajos resultaron insuficientes para el objetivo buscado de medir la inteligencia, debido a que el salto inferencial que existía entre las manifestaciones observables que Galton consideró (pruebas sensorio-motoras) y los atributos psicológicos que pretendía medir con ellas (inteligencia) era demasiado grande; en efecto, los resultados de un test a otro, revelaban fuertes divergencias en una misma persona y apenas se correlacionaban entre sí; por otra parte no se observaron relaciones sustanciales de los valores obtenidos en los tests con el éxito en los estudios.

Posteriormente, en 1895, Binet y Henri publican un artículo en el que critican la práctica totalidad de tests existentes, acusándolos de ser en gran medida sensoriales y de concentrarse en

aptitudes especializadas sencillas. En colaboración con Simon, (Binet y Simón, 1905) introdujo en su escala tareas de carácter más cognoscitivos dirigidas a evaluar procesos superiores como el juicio, la comprensión y el razonamiento, es decir, funciones no sensoriales o aspectos del comportamiento más próximos al concepto de inteligencia que pretendía medir, ampliando así el marco de trabajo de esta perspectiva. La perdida de precisión en la medida de estas características por razón de su mayor nivel de complejidad quedaría compensada con las mayores diferencias en tales variables. Esta escala ha sufrido varias revisiones, algunas de ellas realizadas por el mismo Binet (1908, 1911) y otras no, siendo la más conocida la llevada a cabo por Terman (1916) en la universidad de Standford, donde se utilizó por vez primera el cociente de inteligencia -previamente formulado por Stern- como puntuación en la prueba.

Como señala Navas (1997), *"las escalas de inteligencia de Binet tuvieron un gran impacto ya que, a diferencia de los anteriores tests sensoriales y motores, sí constituían un buen instrumento predictivo en la educación, la clínica y, con algunas modificaciones, en la industria. La adaptación americana de la escala Binet-Simon supuso una auténtica inyección al campo de los tests mentales, donde era muy patente el desencanto producido por la incapacidad mostrada por los tests de Cattell para medir habilidades cognitivas de orden superior"*. (p.35)

En definitiva, aunque estos primeros trabajos resultaron insuficientes, en ellos se establece la idea primigenia que subyace a todos los métodos y teoría de los tests psicométricos que se han desarrollado hasta la actualidad, esto es, la posibilidad de medir indirectamente el "rasgo latente" a través de indicadores empíricos que se consideran manifestaciones del aquel; como señala Yela (1996) un test es *"un reactivo que aplicado a un sujeto revela y da testimonio de la índole o grado de su instrucción, aptitud o manera de ser"* (p. 249); *"la consideración de la "latencia del rasgo" es común a todos los métodos y teorías de los tests psicométricos puesto que es su propia problemática la que ha generado su existencia"* Santisteban (1990, p. 3). Ello ligado a los errores inherentes a la propia situación de medida dado que las medidas obtenidas a partir de las muestras de conducta que forman los tests están afectadas por errores de muestreo, de manera que nunca son exactas. La descripción, estimación y minimización de dichos errores es el principal factor que justifica la necesidad de teorías especializadas en los tests (Crocker y Algina, 1986; Martínez-Arias, 1995).

En definitiva, el desarrollo alcanzado por los tests hacia finales de siglo pasado, provocó el interés por solucionar los problemas teóricos y metodológicos que se derivaron e hizo que fuese necesaria la elaboración de un marco teórico para justificar teórica y formalmente esta manera de medir, es decir, el nacimiento de la Teoría de los Tests.

2.2. Principales teorías de los tests.

En general, el carácter no observable de muchas de las variables psicológica objeto de estudio en nuestra disciplina imposibilita medir directamente a las mismas y obliga a establecer modelos que justifiquen procedimientos que permitan medirlas indirectamente mediante

estimaciones o inferencias. Entre los pioneros que inicialmente contribuyeron significativamente al desarrollo de modelos teóricos, cabe citar, entre otros, a Spearman (1904b), Terman (1916), Thorndike (1904) y Thurstone (1927). Thorndike publica en 1904 el primer libro sobre teoría de los tests con el título *An introduction to the theory of mental and social measurement*. Los trabajos de Spearman y Thurstone han incidido de manera especial en los planteamientos de la medición en psicología. De hecho, como señala Muñiz, el nacimiento formal de la Teoría de los Tests puede ubicarse en los primeros trabajos de Spearman realizados en la primera década en los que da soporte teórico a la integración de las diferentes pruebas de que constan los tests, así como a las puntuaciones que se obtienen con ellas estableciendo los fundamentos de la Teoría Clásica de los Test.

En resumen, de lo anterior, podemos establecer que una teoría de tests es una teoría que proporciona modelos para las puntuaciones de los tests; por tanto, el problema central de una teoría de los tests es la relación que existe entre el nivel del sujeto en la variable no observable que se desea estudiar y su puntuación observada en el test, es decir, el objetivo de cualquier teoría de tests es realizar inferencias sobre el nivel en que los sujetos poseen la característica o rasgo inobservable que mide el test, a partir de las respuestas que éstos han dado a los elementos que conforman el mismo (Muñiz, 1992; Martínez-Arias, 1995). Por consiguiente, para medir o estimar las características latentes de los sujetos es necesario relacionar éstas con la actuación observable en una prueba y esta relación debe de ser adecuadamente descrita por una función matemática que de cuenta del error de medida inherente a toda medición psicológica (estimación del error) y proporcione una estimación del rasgo o característica evaluada (estimación de la característica de interés), esto es, el modelo psicométrico.

Las distintas teorías de tests difieren justamente en la función que utilizan para relacionar la actuación observable en el test con el nivel del sujeto en la variable inobservable. La elección de un tipo de modelo u otro depende de los planteamientos hechos en la elección y formulación de las hipótesis previas, de las condiciones y de los objetivos del test (Santisteban, 1990), es decir, los modelos introducen ciertos supuestos que, de demostrarse que son correctos, validan las inferencias hechas a partir de las puntuaciones.

Así pues, de acuerdo con -Allen y Yen (1979), puede establecerse que una Teoría de los Tests es una representación simbólica de los factores que influyen en las puntuaciones obtenidas con los tests, que está definida por una serie de supuestos en los que se basa.

Desde el trabajo pionero de Spearman, han sido muchos los modelos propuestos bajo la Teoría de los Tests. De entre ellos, Meliá (1990b) distingue dos grandes grupos- la Teoría Clásica de los Tests (TCT) y la Teoría de la Respuesta a los Items (TRI). La TCT tiene sus orígenes en diferentes trabajos de Spearman (1904, 1907, 1910, 1913), (el número 48 del año 1995 de la revista *British Journal of Mathematical and Statistical Psychology* está dedicado a las contribuciones de Spearman a la psicometría. Su relación con la TCT está descrita en el trabajo de Levy, 1995), sistematizados posteriormente por Gulliksen (1950) y por Guilford (1954). Del

modelo clásico se han propuesto numerosas variantes, como por ejemplo el modelo clásico expandido (Bock y Wood, 1971), el modelo de Cureton (1971) o el modelo de Sutcliffe (1965).

La extensión de carácter más global de la TCT la proporciona la Teoría de la Generalizabilidad (TG) (Cronbach, Rajaratnam y Gleser, 1963; Gleser, Cronbach y Rajaratnam, 1965; Cronbach, Gleser, Nanda y Rajaratnam, 1972). Surge como alternativa a la concepción clásica de la fiabilidad y ha alcanzado tal relevancia que se ha ganado un tratamiento equivalente al de la TCT y la TRI en muchos textos (por ejemplo, García Cueto, 1993; Martínez-Arias, 1995, Suen, 1990).

También se han propuesto modelos de supuestos más restrictivos que los de la TCT. Estos modelos son conocidos con el calificativo de Modelos de la Teoría Fuerte de la Puntuación Verdadera, en contraposición a la TCT que en virtud del tipo de supuestos sobre los que se sustenta, también es conocida como la Teoría Débil de la Puntuación Verdadera (véase Keats, 1997).

En cuanto a la Teoría de la Respuesta a los Items, ésta supone un cambio de planteamientos respecto a la TCT, aunque ambas teorías no pueden considerarse contrapuestas sino más bien complementarias (Muñiz, 1997b). Tanto la TCT como la TRI presuponen la presencia de un rasgo subyacente a las respuestas de los sujetos a los ítems del test, y ambas especifican una relación entre esas puntuaciones empíricas y el rasgo latente que, se supone, es el responsable de las puntuaciones. La principal diferencia entre ambas teorías se encuentra en el hecho de que la TCT se establece una relación lineal entre la puntuación del sujeto y su valor en el rasgo medido, mientras en que la TRI esta relación no es lineal (Santisteban, 1990).

El término Teoría de la Respuesta a los Items se debe a Lord (1980), y bajo esa denominación se encuentran un conjunto de modelos, que según Hambleton y Swaminathan (1985) tienen sus orígenes en trabajos publicados en las décadas de los años 30 y 40 (Brodgen, 1946; Ferguson, 1942, Richardson, 1936; Tucker, 1946; Lawley, 1943, 1944). Bock (1997) incluso los relaciona con los modelos de escalamiento de Thurstone, pero que se desarrollaron fundamentalmente a partir de las publicaciones de Lord en la década de los 50 (Lord, 1952, 1953a,1953b), principalmente de su tesis doctoral titulada A theory of tests scores publicada en un número monográfico de la revista Psicometrika (Lord, 1952), en el que presenta un modelo de ojiva normal de dos parámetros para representar la relación entre la habilidad del sujeto y la probabilidad de responder correctamente a los ítems. En la década siguiente se desarrollan modelos logísticos más tratables en un nivel matemático que los modelos de ojiva normal. Rasch (1960), presenta un modelo logística de un parámetro, y Birnbaum (1968) propone en la obra de Lord y Novick (1968) *Statistical theories of mental test scores,* otros dos modelos de dos y tres parámetros. Estos tres modelos logísticos de uno, dos y tres parámetros, constituyen actualmente el tronco básico de la TRI. Desde entonces se han propuesto otros muchos modelos, revisados en extenso por van der Linden y Hambleton (1997), no ya para respuestas de tipo dicotómico como los anteriores, sino para respuestas politómicas, como por ejemplo el de respuesta nominal

(Bock, 1972), el de respuesta graduada (Samejima, 1969) o el de crédito parcial (Masters, 1982), y también para respuesta continua, como el de Samejima (1972). También se han propuesto modelos multidimensionales que en la actualidad están recibiendo mucha atención (Maydeu, 1996; Reckase, 1997). Pueden encontrarse más detalles sobre el desarrollo de los modelos de la TRI en Lord, 1980; Thorndike, 1982; Hulin, Drasgow y Parsons, 1983; Baker, 1985; Hambleton y Swarninathan, 1985; Crocker y Algina, 1986; Goldstein y Wood, 1989; Hambleton, 1990; Meliá, 1990b; Muñiz, 1990; Hambleton, Swaminathan y Rogers, 1991; Hambleton y Zaal, 1991/1994; Muñiz y Hambleton, 1992; López Pina, 1995; Martínez Arias, 1995; Muñiz, 1996a; Bock, 1997; y van der Linden y Hambleton (1997b).

Como reacción a los planteamientos de la TCT, principalmente en lo que se refiere a la manera de interpretar las puntuaciones a partir de normas de grupo, surge en la década de los años 60 (Glaser, 1963) un nuevo tipo de tests denominados Test Referidos al Criterio (TRC) (véase Hambleton, 1997a). Con los tests referidos al criterio se cambia el énfasis de la referencia a la norma a la referencia a un criterio en la interpretación de las puntuaciones de los tests. Su justificación teórica puede hacerse tanto desde la TCT, como desde la TG o la TRI.

En resumen, podemos establecer que las principales teorías de tests que han surgido en el campo de la psicometría son:

- Teoría clásica de los tests
- Teoría de la generalizabilidad
- Teoría de respuesta a los ítems.

A continuación describiremos la teoría clásica de los tests (dejando la teoría de la generalizabilidad y la teoría de respuesta a los ítems para los dos próximos capítulos). En cada una de ellas, señalaremos sus características más relevantes, una breve reseña de su evolución histórica, los planteamientos básicos y fundamentales que las caracterizan y las principales ventajas y limitaciones que presentan.

Puede encontrarse más información al respecto en cualquier manual de psicometría. En la revista *Educational Measurement.- Issues and Practice* se dedicó un número monográfico (1997, Vol. 16, Nº 4) a la historia de la psicometría, en el que se pueden encontrar una revisión histórica de la TCT (Traub, 1997), de la TG (Brennan, 1997) y de la TRI (Bock, 1997). De especial interés, por su enfoque crítico, son también los trabajos de Lumsden (1976) y el más reciente de Blinhom (1997).

2.3. Teoría Clásica de los Tests (TCT).

2.3.1. Introducción a la Teoría Clásica de los Tests (TCT).

El primer modelo en el ámbito psicológico que aborda el problema del error de las medidas realizadas mediante la aplicación de un test, fue el presentado por Spearman en 1904, donde planteó el que ya viene a ser el clásico “Modelo Lineal de Puntuaciones”, denominado

"Teoría débil de la Puntuaciones Verdadera" o simplemente "Teoría Clásica de los Tests". Aunque este modelo como tal, nunca formó parte de un texto completo e integrado por parte de su progenitor, sino que su formulación inicial se debe a una serie de artículos que publicó a principios del siglo (1904, 1907, 1910, 1913), sus planteamiento fueron elaborados y ampliados con las aportaciones de un gran número de investigadores durante la primera mitad del siglo XX, siendo expuesta por primera vez de forma completa en la obra de Guilford (1936) *Psychometric Methods;* en 1950, en su obra *Theory of Mental Tests*, Gulliseck revisa y sistematiza todo el *corpus* teórico de conocimiento desarrollado hasta el momento en torno a esta teoría. Con posterioridad, Lord y Novick (1968), en su ya clásica obra *Statistical Theory of Mental Test Scores,* presentan una reformulación del modelo en términos de la teoría estadística.

A pesar de las numerosas limitaciones que se le atribuyen y del desarrollo de otras aproximaciones alternativas o complementarias, la TCT sigue estando vigente y continúan apareciendo manuales en que se trata este modelo y trabajos en las principales revistas, aunque en número muy escaso comparado con otras teorías de los tests (v.gr., Allen y Yen, 1979; de Guijter y van der Kamp, 1983; Croker y Algina, 1986; Nunnally y Bernstein, 1987; Santisteban, 1990; Muñiz, 1992; García Cueto, 1993; Meliá, 1993; Levy, 1995; Martínez-Arias, 1995; Traub, 1997; Zimmerman y Williams, 1997). En definitiva, como señala Santisteban (1990) *"Aún cuando posteriormente se han adoptado nuevas teorías, la teoría basada en el modelo de Spearman ... sigue siendo influyente en nuestro tiempo, y así lo reconocen autores especializados como lo son entre otros Weiss y Davison, que lo hacen en el Annual Review of Psychology en 1981."* (p.26).

La tesis central de Spearman era formalizar un modelo estadístico que pudiese tener en cuenta la estimación de los errores de medida inherente a todo proceso de medición. De este modo se posibilitaría una fundamentación adecuada para las puntuación de los tests en el campo de la psicología

Para elaborar la teoría, Spearman asumió que la puntuación empírica de un sujeto en un test (***X***), es susceptible de descomponer en dos partes o componentes aditivos que directamente no se pueden observar; por un lado la puntuación verdadera del sujeto (***V***) en el rasgo tal como lo mide el test, y por otro lado el error aleatorio de medida (***e***) que inevitablemente va asociado a las puntuaciones de los tests; formalmente se expresa como:

$$\boldsymbol{X = V + e} \qquad (2.1)$$

Jöreskog (1973) expresa la hipótesis fundamental del modelo clásico en términos de un modelo de análisis factorial con un factor común como sigue:

$$\boldsymbol{x = \Lambda\xi + \delta} \qquad (2.2)$$

donde ***x*** es un vector columna de dimensiones (***q*** x 1) compuesto por la ***q*** variables observadas o ítems; ***Λ*** es una matriz de dimensiones (***q*** x 1) compuesta por las saturaciones factoriales que relacionan las variables observadas ***x*** con la variable latente (ξ) o factor común; ξ es un vector

que contiene la variable latente o factor común (en nuestro caso, las puntuaciones verdaderas de los sujetos en esas variables o ítems); y $\boldsymbol{\delta}$ es un vector columna de dimensiones ($\boldsymbol{q}$ x 1) compuesto por los factores únicos (en nuestro caso, la parte de error aleatorio de las mediciones empíricas; aunque estos factores únicos pueden ser considerado como combinación lineal de dos componentes independientes entre sí: un componente específico asociado a cada una de las variables observadas (***s***) y un componente de error aleatorio (***e***)).

Es decir, el modelo que propone Spearman, es el modelo lineal clásico, $\boldsymbol{f}(\mathbf{Y}) = \mathbf{a}\,\boldsymbol{X} + \mathbf{b}$. Con este modelo y los supuestos inherentes al mismo, la teoría clásica desarrolla todo un conjunto de deducciones que permiten estimar la cantidad de error que afecta a las puntuaciones de los tests. (véase la figura 2.1.).

Teoría Clásica de los Tests (Spearman):

$$\boldsymbol{X} = \boldsymbol{V} + \boldsymbol{error}$$

Propone el modelo lineal:

$$\boldsymbol{x} = \boldsymbol{\lambda}\,\boldsymbol{\xi} + \boldsymbol{\delta}$$

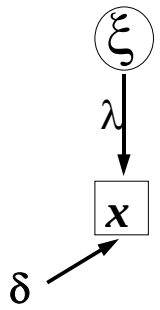


Figura 2.1.- Expresión y representación del modelo clásico.

En el sentido de los párrafos anteriores, la puntuación empírica, (***X***), que se obtiene para un sujeto cuando se le aplica un test en un momento dado, razonablemente no tiene porque coincidir exactamente con su verdadera puntuación, (***V***), pues en ese momento el sujeto puede estar afectado por múltiples factores no controlados que inciden en su conducta, (***e***). Es por esto que la aceptación generalizada de este modelo proviene del hecho elemental de que los errores no son observables directamente, y que la utilización del mismo posibilita la descomposición de la medición empírica (***X***) en sus dos componentes aditivos (***V***) y (***e***) de un modo simple y parsimonioso, mediante el uso de los tres supuestos básicos que se asumen para el modelo.

2.3.2. Supuestos fundamentales de la teoría clásica de los tests.

En primer lugar se define el concepto de puntuación verdadera de un sujeto como la esperanza matemática de la puntuación empírica observada en infinitos ensayos de medidas independientes realizados con ese individuo, y, considerando el error de medida como un componente aleatorio que simplemente se suma a la puntuación verdadera del sujeto en el test, es decir, la puntuación en el test libre del componente de error. En este sentido el primer supuesto

establece que la puntuación verdadera (***V***) coincide con el valor esperado de la puntuación empírica; por tanto la puntuación verdadera es un concepto matemático y como tal puede ser estimado, esto es,

$$V = E(X) \tag{2.3}$$

Los siguientes supuestos hacen referencia a la naturaleza del error de medida, y a las relaciones que se esperan entre el error de medida y la puntuación verdadera de los sujetos, y, entre los errores de medida de diferentes mediciones, es decir, el error de medida se considera como una variable aleatoria que sigue una distribución normal con media cero y varianza σ^2_e y las varianzas de los errores son iguales cualquiera que sea la puntuación verdadera a la que vayan asociados, es decir, el modelo es homocedástico. Formalmente:

Naturaleza del error de medida:

$$E(e) = 0 \tag{2.4}$$

$$Var(e_i) = \sigma^2_{e_i} = \sigma^2_{e_i} = Var(e_j) \text{ para todo } i,j \tag{2.5}$$

Relación entre errores y puntuaciones verdaderas:

- No existe correlación entre las puntuaciones verdadera de los sujetos de un test y sus respectivos errores de medida:

$$\rho_{ve} = 0 \tag{2.6}$$

Relación entre errores de medida

- No existe correlación entre los errores de medida de dos medidas diferentes, es decir, los errores de medida de los sujetos en un test no correlaciona con sus errores de medida en otro test distinto, ni siquiera en otra aplicación del mismo test a los mismos sujetos, :

$$\rho_{e_i e_j} = 0 \tag{2.7}$$

Estos supuestos no pueden ser comprobados empíricamente, es decir, no permiten deducir la cantidad de error que afecta a una determinada puntuación en un test. En este sentido, como señalan Hambleton y van der Linden (1982), se trata de tautologías; sin embargo, a partir del modelo y de sus supuestos básicos (véase Allen y Yen, 1979) se pueden extraer deductivamente, una serie de relaciones e índices a través de los cuales sí se pueden obtener estimaciones del error cometido en las puntuaciones observadas en un test, facilitando así la aplicación práctica del mismo. Las principales deducciones que se derivan de los supuestos son las siguientes:

- El valor esperado de la puntuación verdadera es igual al valor esperado de la puntuación empírica,

$$E(V) = E(X) \tag{2.8}$$

- La ecuación de regresión de la puntuación empírica sobre la puntuación verdadera es la ecuación lineal que pasa por el origen y que tiene el valor unidad como pendiente de la recta. La regresión que se obtiene es la de la variable X sobre cada uno de los valores verdaderos V= v_k, para k= 1,...,N. Formalmente se puede expresar como sigue:

$$E(X \mid V = v_k) = v_k \qquad (2.9)$$

- La varianza de las puntuaciones empíricas en un test es igual a la suma de la varianza de las puntuaciones verdadera más la varianza de los errores de medida:

$$\sigma_x^2 = \sigma_v^2 + \sigma_e^2 \qquad (2.10)$$

- La covarianza entre las puntuaciones empíricas y las puntuaciones verdaderas es igual a la varianza de las puntuaciones verdaderas:

$$COV(X,V) = VAR(V) = \sigma_v^2 \qquad (2.11)$$

- El cuadrado del coeficiente de correlación entre las puntuaciones empíricas y sus correspondientes puntuaciones verdaderas es igual a la razón de la varianza de las puntuaciones verdaderas con respecto a la varianza de las empíricas:

$$\rho_{xv}^2 = \frac{\sigma_v^2}{\sigma_x^2} \qquad (2.12)$$

- La covarianza entre las puntuaciones empíricas y los errores de medidas es igual a la varianza de los errores:

$$COV(X,e) = VAR(e) = \sigma_e^2 \qquad (2.13)$$

- El cuadrado del coeficiente de correlación entre las puntuaciones empíricas y sus correspondientes errores de medida es igual a la razón de la varianza de los errores con respecto a la varianza de las puntuaciones empíricas:

$$\rho_{xe}^2 = \frac{\sigma_e^2}{\sigma_x^2} \qquad (2.14)$$

- El cuadrado del coeficiente de correlación entre las puntuaciones empíricas y sus correspondientes puntuaciones verdaderas es igual a 1 menos el cuadrado del coeficiente de correlación entre las puntuaciones empíricas y sus correspondientes errores de medida:

$$\rho_{xv}^2 = 1 - \rho_{xe}^2 \qquad (2.15)$$

Llegados a este punto, el modelo así formulado, con sus supuestos y derivaciones, no tienen relevancia práctica, es decir, resultan inoperante, porque todos ellos contienen de una u otra forma elementos no observables empíricamente. Por esta razón Spearman introduce el concepto de tests paralelos, que le permitirá operar empíricamente con las puntuaciones obtenidas directamente por los sujetos en los tests.

2.3.3. Concepto de test paralelos: el modelo de formas paralelas y sus variantes.

Spearman asume que se puede construir dos o más test que midan lo mismo aunque con diferentes ítems, es decir, dos formas equivalentes del mismo test. En este sentido dos conjuntos de puntuaciones ***X*** y ***X'***, se dice que son equivalentes o paralelas si ambas tienen la misma puntuación verdadera, ***X=V+e*** y ***X'=V+e'***, y además, ambas poseen la misma varianza de error, ***Var(e) = Var(e'),*** (véase la figura 2.2.). A partir de estos supuesto, en el caso de que se cumplan, se puede establecer de modo simple e intuitivo que para k tests paralelos, desde el punto de vista de sus valores paramétricos, las medias poblaciones son idénticas, cada test paralelo presenta la misma varianza poblacional y las intercorrelaciones entre cada par de ellos son iguales. Es decir,

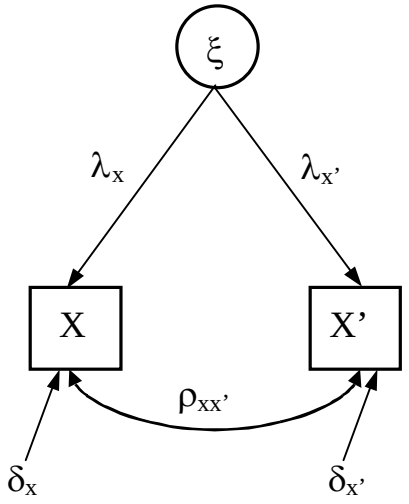
$$\mu_1 = \mu_2 = \ldots = \mu_k$$
$$\sigma^2_{(X_1)} = \sigma^2_{(X_2)} = \ldots = \sigma^2_{(X_k)} \qquad (2.16)$$
$$\rho_{(X_1,X_2)} = \rho_{(X_1,X_3)} = \ldots = \rho_{(X_j,X_k)}$$

Teoría Clásica de los Tests (Spearman): medidas paralelas

$$X = V + e \quad y \quad X' = V + e'$$

$$\sigma^2_e = \sigma^2_{e'} \qquad \rho_{xx'} = \frac{\sigma^2_v}{\sigma^2_x}$$

Modelo de formas paralelas:



$$x = \lambda_x \xi + \delta_x \qquad x' = \lambda_{x'} \xi + \delta_{x'}$$

$$\lambda_x = \lambda_{x'} \qquad \theta^2_{\delta_x} = \theta^2_{\delta_{x'}} \qquad \rho_{xx'} = \lambda_x \cdot \lambda_{x'}$$

Figura 2.2.- Expresión formal y representación gráfica del modelo de formas paralelas.

Estos supuestos, que se refieren por fin a las puntuaciones empíricas, sí se pueden comprobar y dado que la Teoría Clásica de los Tests asume el modelo probabilístico y que los estadísticos obtenidos en muestras suficientemente amplias constituyen buenos estimadores de los parámetros poblacionales, las deducciones expuestas pueden ser fácilmente sometidas a contrastación. De las condiciones de paralelismo se deducen múltiples consecuencias, entre las cuales, la más importante es que la correlación entre dos medidas paralelas expresa la proporción de varianza en puntuaciones empíricas que es debida a la varianza de las puntuaciones verdaderas, esto es, $\rho_{xx'} = (\sigma^2_v / \sigma^2_x)$. De este modo, el coeficiente de correlación o coeficiente de fiabilidad se convierte en la piedra angular del modelo clásico; precisamente, el coeficiente de fiabilidad de un test -expresión operativa de uno de los dos criterios métricos de calidad que hay que exigir a todo instrumento de medida-, se define como la correlación entre dos formas paralelas de un test y como señala Lloret (1999) *"responde a la necesidad que había dado lugar a la formulación del modelo: estimar el efecto del error de medida en las puntuaciones de los tests"*, un efecto hasta entonces imposible de estimar.

Ahora, la deducción de la varianza del error es inmediata, y su desviación típica resulta fácilmente deducible:

$$\sigma_e^2 = \sigma_x^2(1-\rho_{xx'}) \Rightarrow \sigma_e = \sigma_x\sqrt{1-\rho_{xx'}} \qquad (2.17)$$

Sin embargo, aunque el paralelismo es el concepto fundamental, podemos decir que ésta es la gran restricción que lleva aparejada el modelo; la contrastación empírica de las condiciones de paralelismo entre dos tests fueron difíciles de comprobar y no se han resuelto satisfactoriamente hasta 1946, año en que Wilks propuso el estadístico LMVC que permitió comprobar simultáneamente si las medias y varianzas poblacionales de dos series de medidas podían ser consideradas iguales, condición necesaria y suficiente para sostener la igualdad de las puntuaciones verdaderas y de la varianza de error.

Posteriormente, una alternativa a las técnicas tradicionales de inferencia estadística, para determinar el paralelismo entre dos test, usada muy frecuentemente lo constituye el uso del análisis factorial confirmatorio y los modelos de ecuaciones estructurales (Gómez-Benito, 1996). Como indica Martínez-Arias (1995), aún en el difícil caso de que con dos tests se obtengan puntuaciones verdaderas iguales, resultaría difícil que sus varianzas de error sean idénticas. Estas situaciones quedan contempladas en las formulaciones de Lord y Novick (1968) y Jöreskog (1968) que proponen tres formas de paralelismo en que se relajan algunos de los supuestos originales del modelo de formas paralelas:

- las medidas *tau-equivalentes* (tau es la letra griega que representa a la puntuación verdadera, true en inglés) (Lord y Novick, 1968), asumen igualdad en las puntuaciones verdaderas y diferencias en las varianzas de error, es decir, se refiere a tests con los que se obtienen las mismas puntuaciones verdaderas de los sujetos pero con diferentes varianzas de error,

- las medidas *esencialmente tau-equivalentes* (Lord y Novick, 1968), asumen puntuaciones verdaderas diferenciadas por un valor constante, es decir, diferentes aunque perfectamente correlacionadas. Las varianzas de error pueden ser iguales o diferentes.
- las medidas *congenéricas* (Jöreskog, 1968), en las que además de varianzas de error distintas se caracterizan por que las dos puntuaciones verdaderas obtenidas con los tests son transformaciones lineales una de la otra. En este modelo de paralelismo cada medida refleja la misma puntuación verdadera, pero en distintos grados y con diferentes promedios de error de medición.

Feldt y Brennan, 1989, presentan una forma alternativa de paralelismo que denominan medidas *congenéricas multifactoriales*, donde la puntuación empírica obtenida por un sujeto en un test es una combinación ponderada de una serie de puntuaciones verdaderas en diferentes rasgos, y esa ponderación puede ser diferente en un test distinto.

Con la cuantificación o estimación del error de medida estamos en condiciones de estimar el intervalo de confianza del nivel de medida en que un sujeto determinado posee la característica o rasgo que mide el test.

2.3.4. Estimación de la puntuación verdadera.

Como ya se ha señalado, el nivel real del sujeto en la característica de interés, puntuación verdadera (***V***), es la media de los valores que se obtendrían de forma empírica en caso de administrar el mismo test al sujeto en idénticas condiciones de medida un número infinito de veces, es decir, el valor esperado o valor del parámetro media de la variable medida, y, la relación entre el comportamiento observable en el test -***X***- y el nivel del sujeto en la variable no observable -V- es una relación lineal. La estimación del intervalo de confianza del nivel real en que un sujeto determinado posee la característica o rasgo que mide el test obedece a la siguiente formula general:

$$P(\hat{V} - E.\text{máx.} \leq V \leq \hat{V} + E.\text{máx.}) \leq \alpha \qquad (2.18)$$

Para determinar el valor de ***V*** y del error máximo (***E***.máx.) que podemos asumir en la estimación del valor verdadero (***V***) se dispone de tres estrategias:

- Estimación mediante el teorema de Chevychev.
- Estimación basada en la distribución normal de los errores.
- Estimación según el modelo de regresión.

3.3.4.1. Estimación mediante el teorema de Chevychev.

$$\hat{V} = X$$

$$E.max = k\sigma_e = \sigma_e\sqrt{\frac{1}{\alpha}} \qquad (2.19)$$

Donde,

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{xx'}} \qquad (2.20)$$

Por tanto,

$$P\left(X - \sigma_e \sqrt{\frac{1}{\alpha}} \leq V \leq X + \sigma_e \sqrt{\frac{1}{\alpha}} \right) \leq \alpha \qquad (2.21)$$

3.3.4.2. Estimación basada en la distribución normal de los errores.

$$\hat{V} = X$$
$$E.max = z_c \sigma_e \qquad (2.22)$$

Donde,

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{xx'}} \qquad (2.23)$$

Por tanto,

$$P\left(X - z_c \sigma_e \leq V \leq X + z_c \sigma_e\right) \leq \alpha \qquad (2.24)$$

3.3.4.3. Estimación según el modelo de regresión.

$$\hat{V} = V' = \rho_{xx'}\left(X - \mu_x\right) + \mu_x$$
$$E.max = z_c \sigma_{v.x} \qquad (2.25)$$

Donde,

$$\sigma_{v.x} = \sigma_e \sqrt{\rho_{xx'}} \qquad (2.26)$$

Por tanto,

$$P(\hat{V} - z_c \sigma_{v.x} \leq V \leq \hat{V} + z_c \sigma_{v.x}) \leq \alpha \qquad (2.27)$$

Por otro lado, como se ha señalado, según la teoría clásica de los tests, la puntuación empírica que obtiene un sujeto cuando se le administra un test -***X***- es función del nivel real o verdadero en que el sujeto posee la característica o rasgo que está evaluando dicho test (***V***), y el error de medida que siempre se introduce en cualquier proceso de medición, -***e***-. Por tanto, si podemos estimar la puntuación verdadera de un sujeto, también podemos hacer lo mismo respecto del error de medida.

2.3.5. Estimación del error de medida.

El error de medida podemos considerarlo como la diferencia entre la puntuación empírica u observada y la puntuación verdadera de la característica que se le ha medido a un sujeto mediante un test, esto es,

$$e = X - V \qquad (2.28)$$

Por otro lado, en la estimación del valor verdadero (V) de la característica a medir, se comete un error de estimación inherente a los procedimientos de estimación. Este tipo de error lo denominamos error de estimación de la puntuación verdadera. Formalmente se expresarsa como,

$$E = \hat{V} - V \qquad (2.29)$$

Donde la puntuación verdadera estimada es la puntuación verdadera pronosticada a partir de la puntuación empírica mediante el modelo de regresión, esto es,

$$\hat{V} = \rho_{XX'}(X - \mu_x) + \mu_x \qquad (2.30)$$

Por tanto, es conveniente diferenciar claramente entre el error típico de medida y el error típico de estimación de la puntuación verdadera:

Error típico de medida:

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{xx'}} \qquad (2.31)$$

Error típico de estimación de la puntuación verdadera:

$$\sigma_{v.x} = \sigma_e \sqrt{\rho_{xx'}} \qquad (2.32)$$

Continuando con la descripción del modelo de la TCT, éste ha sido el modelo dominante en la teoría de los tests durante casi todo el siglo XX y, aun hoy día, sigue teniendo gran influencia en la formación de los profesionales de la psicología y otras ciencias afines (psicopedagogía, pedagogía, sociología...) y se le otorga un peso importante en los planes de estudios y manuales de estas disciplinas. Es notable su vigencia en el campo de la práctica de la evaluación psicológica y educativa, y, aunque presenta grandes ventajas también podemos atribuirle grandes debilidades.

2.3.6. Ventajas y limitaciones de la TCT.

Como señala Muñiz (1992), lo que le ha proporcionado una larga vida a la TCT son su sencillez matemática y enjundia psicológica. Crocker y Algina (1986) apuntan que el éxito de la TCT se debe a que proporciona solución a una amplia gama de problemas de medida y a que sus supuestos son mínimos. En definitiva, la sencillez, claridad y flexibilidad de sus conceptos, junto a la simplicidad de sus supuestos y procedimientos han posibilitado que pueda ser aplicada a muchas situaciones en las que no tienen cabida modelos con supuestos más restrictivos.

Ahora bien, sus ventajas conllevan muchos problemas y algunos de ellas son ciertamente importantes. Entre los más relevantes podemos destacar los siguientes:

Limitaciones relacionadas con los supuestos del modelo.

Los supuestos del modelo no pueden ser contrastados empíricamente; por definición tienen un carácter tautológico, es decir no pueden ser evaluados (Lord y Novick, 1968; Warm, 1978; Hambleton y van der Linden, 1982).

Limitaciones relacionadas con el concepto de tests paralelos.

La equivalencia de las medidas paralelas, en la práctica, es difícil de conseguir. Este supuesto que es clave en la TCT resulta muy restrictivo y rara vez se cumple (Lord y Novick, 1968; Hambleton y van der Linden, 1982; Hambleton, 1989; Muñiz 1996).

Limitaciones relacionadas con los parámetros del modelo.

Los parámetros del modelo no son invariantes, es decir, las puntuaciones de los sujetos dependen de los ítems del test y éstos a su vez dependen de la muestra de sujetos a los que se ha aplicado el test. Como señalan Muñiz y Hambleton, 1992, "*si se aspira a una medición rigurosa y científica, resulta difícil justificar que las mediciones estén en función del instrumento utilizado"* (p.40).

Limitaciones relacionadas con el concepto de fiabilidad.

La fiabilidad del test es un concepto central en la TCT, y sin embargo no es posible definirlo ni estimarlo de forma unívoca (Lloret, 1999). La implementación de múltiples procedimientos para estimarla ha producido confusión en los conceptos de consistencia interna, homogeneidad y unidimensionalidad y, en consecuencia, el valor del coeficiente de fiabilidad depende del método de estimación usado. Ello, unido a que la fiabilidad del instrumento de medida depende de la longitud del test y de la variabilidad de las respuestas de los sujetos a los que se les aplica el test, hace que la consideración de la fiabilidad como una propiedad característica del instrumento de medida sea difícil de sostener.

Limitaciones relacionadas con los errores de medida.

El tratamiento que se da a los errores de medidas en la TCT es uno de los aspectos mas discutidos debido a los problemas que se derivan del incumplimiento del carácter completamente aleatorio que asume el modelo. La consideración de homocedasticidad de las varianzas de los errores para todos los niveles de habilidad de los sujetos y la independencia de estos errores respecto a las puntuaciones verdaderas es difícil de asumir, sobre todo, para los valores extremos de la escala. En este sentido, se ha reconocido que el error típico de medida varía con el nivel de habilidad (Mollenkopf, 1949; Thorndike, 1951; Feldt, Steffen y Gupta , 1985).

Otro aspecto discutido se refiere a la independencia de los errores entre dos conjuntos de mediciones, dos aplicaciones de un mismo test o de formas paralelas. Para la evaluar la idoneidad de este supuesto se pueden utilizar los modelos de estructuras de covarianza, (Gómez-

Benito, 1996), valorando el mejor ajuste resultante en función liberar o no la restricción de independencia de los errores.

La consideración simplista de las fuentes de los errores de medida, es otra de las críticas, a mi juicio, más importantes al modelo; el carácter único e indiferenciado del error de medida obliga a incluir dentro del mismo todas las posibles fuentes de error (variaciones individuales, factores situacionales, características del aplicador, variables instrumentales, ..), es decir , el modelo presenta serias dificultados al no poder diferenciar las distintas fuentes de error que afectan a las puntuaciones (Cronbach, Glesser, Nanda y Rajaratnam, 1972).

En resumen, las limitaciones expresadas, junto a otros problemas que presenta el modelo, como son las limitaciones de los tests normativos estandarizados ligadas a su validez externa o generalización restringida de las aplicaciones (dado que las inferencias posibles dependen del grado de representatividad de los sujetos), la ausencia de explicación del comportamiento de los sujetos a nivel de los ítems particulares que conforman el instrumento de medida (van der Linden, 1986; Hambleton, 1989; Hambleton, Swaminathan y Rogers, 1991), las limitaciones del mismo al no proporcionar soluciones adecuadas a temas prácticos como el diseño de tests, los tests adaptativos, el sesgo y la equiparación de puntuaciones (Martínez-Arias, 1995), no han impedido que este modelo siga aún vigente, conviviendo, a nivel de la enseñanza universitaria, con desarrollos más actuales que superan muchas de sus deficiencias (Teoría de la Generalizabilidad y Teoría de Respuestas a los Ítems) y, en el marco profesional, la practica totalidad de las aplicaciones están aún diseñadas desde el modelo clásico. La sencillez y simplicidad que le caracterizan hace que siga siendo atractivo e imprescindible cuando las exigencias de otros modelos impiden su aplicación.

No obstante, sus limitaciones han producido reacciones muy variadas encaminadas a proponer tentativas de solución; muchas de estas propuestas se han traducido en variantes y/o desarrollos de modelos alternativos a la TCT. Entre los más relevantes, la teoría de la Generalizabilidad y la Teoría de Respuesta a los Ítems. A continuación presentamos algunos de estos desarrollos.

2.3.7. Variantes del modelo clásico lineal.

Como señala López-Feal (1986), en líneas generales, las propuestas de alternativas han sido de tres tipos: a) la continuidad, desarrollando nuevos métodos para calcular la fiabilidad, b) la extensión de la teoría mediante su reformulación, c) ruptura y abandono de la teoría. Todas ellas encaminadas a superar las limitaciones del modelo.

Así, nos encontramos con modelos que han sido reformulados desde la insatisfacción con el supuesto de paralelismo que afecta al concepto de fiabilidad. En los párrafos anteriores, hemos expresado los planteamientos de los modelos basados en formas de equivalencia menos restrictivas como las medidas *tau-equivalentes*, *esencialmente equivalentes* y *congenéricas*. Todos ellos asumen la posibilidad de que las puntuaciones puedan tener distinta varianza de error

y se diferencian en el tipo de semejanza que se establece entre las puntuaciones verdaderas. El *modelo muestral de ítems* (Tryon, 1957; Nunnally, 1987) presenta un cambio en el enfoque del concepto paralelismo y propone el supuesto de tests aleatoriamente paralelos. Se trata de un supuesto más débil que considera a los ítems de los tests como muestras aleatorias extraídas de una población universo compuesta por todos los ítems posibles que permiten medir el constructo subyacente (Hontangas, 1997).

Otros modelos, se caracterizan por rechazar el supuesto que considera toda la variación sistemática en las puntuaciones observadas debida sólo a las diferencias individuales en el rasgo medido, admitiendo la existencia de varios tipos de errores, es decir, contemplan que otras fuentes de errores (condiciones de aplicación, característica del test, de los sujetos,..) influyen en las puntuaciones. Así el modelo *clásico expandido* (Bock y Wood, 1971) que propone una descomposición del término clásico de error en dos componentes independientes, el error sesgado o sistemático ($\mathbf{E}_s$) y el error insesgado o residual ($\mathbf{E}_r$), que, aditivamente, junto con la puntuación verdadera conforman la puntuación empírica $\mathbf{X} = \mathbf{V} + \mathbf{E}_s + \mathbf{E}_r$. El modelo asume que la esperanza matemática del error residual es nula y que no hay relación entre la puntuación verdadera y el error sistemático. El modelo *platónico* propuesto por Sutcliffe, 1965, que asume al modelo anterior, $\mathbf{X} = \mathbf{V} + \mathbf{E}_s + \mathbf{E}_r$,pero con el supuesto de que entre la puntuación verdadera y el error sesgado puede haber relación. El modelo de Cureton (1971) es similar a los anteriores pero con el supuesto de que los dos tipos de error son independientes y que tampoco existe relación entre la puntuación verdadera y ninguno de los dos tipos de error. El *modelo clásico operacional* que define la puntuación verdadera incorporando en su delimitación el componente sesgado del error,$\mathbf{V}^* = \mathbf{V} + \mathbf{E}_s$,, asumiendo todos los supuestos del modelo clásico.

Como señala Navas (1997) ninguno de estos modelos representan una alternativa real al modelo clásico, “son modelos, de algún modo, anecdóticos, en el sentido de que son simples reformulaciones del modelo clásico en las que bien se redefine un aspecto, bien se introduce algún parámetro nuevo, pero no suponen, en ningún caso, un planteamiento nuevo o diferente de conceptos básicos en el campo de la medida como, por ejemplo, la fiabilidad” (p.121).

Entre los modelos que se han formulado desde la insatisfacción con el tratamiento que recibe el componente de error en se encuentra la Teoría de la Generalizabilidad (Cronbach, Gleser, Nanda y Rajaratnam, 1972) que abandona el concepto clásico de test centrado en el binomio “test-sujeto”, y el concepto de fiabilidad basado en la idea de correlación entre puntuaciones empírica. La descripción de sus planteamientos básicos se expondrá en el capítulo siguiente.

Otros modelos que se fundamentan en el mismo modelo que la TCT pero con supuestos más exigentes sobre la distribución de sus componentes. Estos modelos se denominan Teoría Fuerte de las Puntuaciones Verdaderas y establecen distribuciones específicas para los errores de medida que pueden comprobarse empíricamente (Lord, 1965).

Este tipo de modelos comparte, en su estructura básica, el modelo y los supuestos de la TCT, a excepción del supuesto de incorrelación entre los errores obtenidos con tests diferentes. Además admiten la posibilidad de que la distribución condicional de los errores respecto a las puntuaciones verdaderas puede ser heteroscedástica. Este supuesto adicional permite distinguir entre modelos que suponen una distribución condicional de los errores de tipo binomial (Keats y Lord, 1962 y Lord, 1965) o de tipo Poisson (Rasch, 1960).

Los modelos de la Teoría Fuerte han recibido poca atención en los textos psicométricos, pero aún así pueden encontrase descritos en Allen y Yen (1979), Lord y Novick (1968) o Santisteban (1990). Sus planteamientos básicos se expondrán en el capítulo siguiente.

Como reacción a las limitaciones de los tests normativos, Glaser y Klaus (1962), han propuesto un conjunto de estrategias y técnicas específicas que, sin constituir una nueva teoría de los tests, pretenden construir y evaluar instrumentos que permitan interpretar las puntuaciones en su sentido absoluto (Medidas Referidas al Criterio) y describir con mayor precisión los conocimientos, habilidades y destrezas de los sujetos en un dominio concreto de contenido. En el siguiente capítulo se desarrollarán sus planteamientos básicos.

Por último encontramos modelos que superan las limitaciones relacionadas con la dependencia de la población y con la ambigüedad de las puntuaciones verdaderas. Son modelos que incorporan explícitamente parámetros relacionados con los sujetos y con los ítems que conforman el test escalando directamente a los sujetos y a los ítems en el rasgo que miden. Estos modelos se engloban bajo la denominación general de Teoría de Respuestas a los Ítems, y, como señala Muñiz (1997), suponen un cambio radical a los planteamientos de la TCT, aunque no llegan a ser teorías contrapuestas sino complementarias. La TRI es la teoría de los tests que dominada en la actualidad y está llamada a sustituir a la TCT.

A continuación, en el capítulo siguiente se desarrollan los aspectos fundamentales en que se basan las principales desarrollas que hemos mencionado en los párrafos anteriores. La TRI, dado la extensión de su desarrollo actual se presentará en un capitulo posterior.