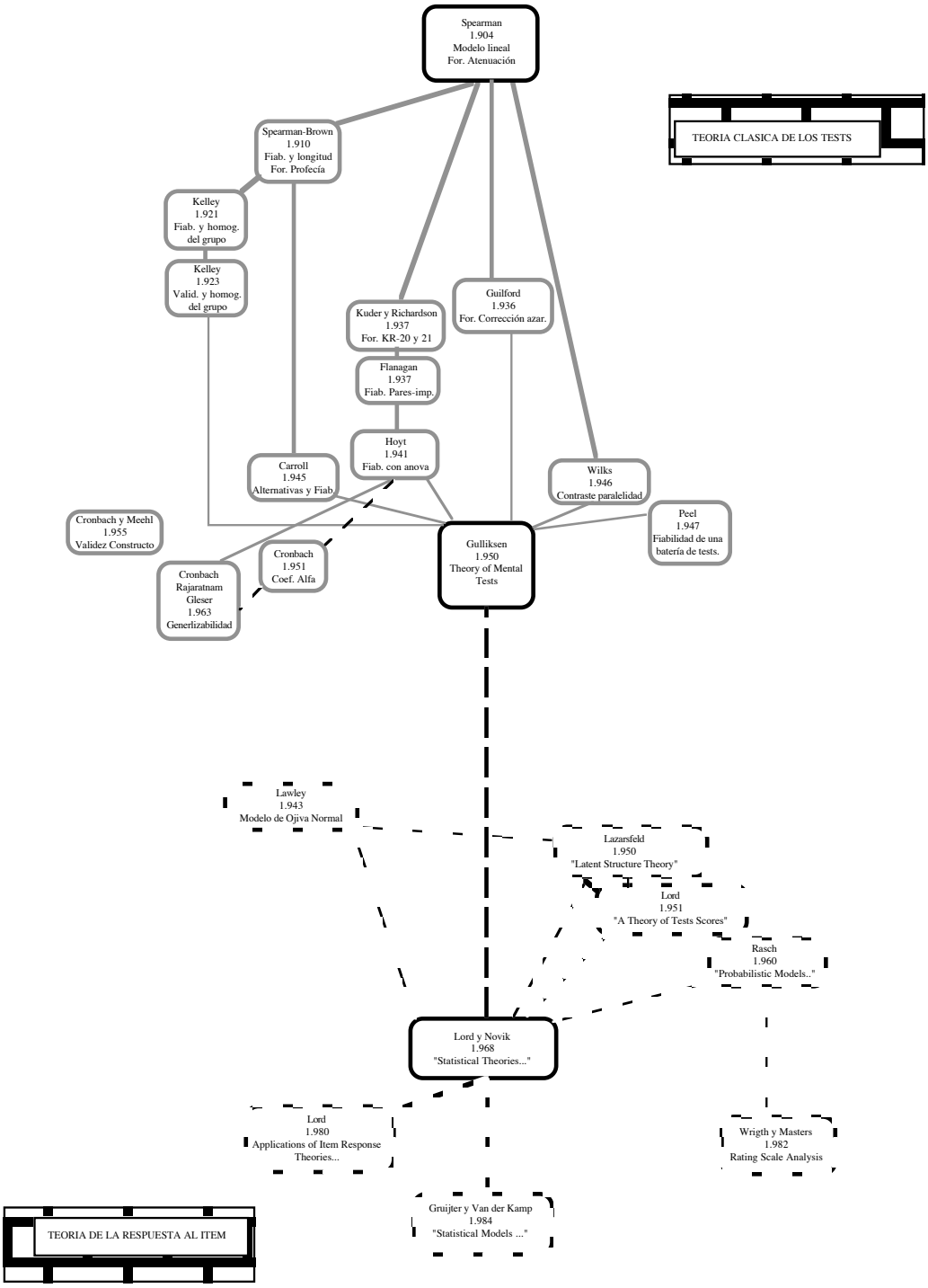
6. TEORIA CLASICA DE LOS TESTS.

1. Nacimiento de la teoría clásica de los tests.

Contenidos del concepto de "teoría clásica". La teoría clásica de los tests, tal como ahora se la conoce, es un conjunto con función de pertenencia no del todo bien establecida que incluye sin duda las aproximaciones esenciales sobre *teoría de la fiabilidad,* incluyendo las aportaciones de Spearman, de Kuder y Richardson, de Wilks, de Kelley, parte del trabajo de Thurstone, etc., y los desarrollos clásicos en *análisis de items, tipificación y validez.* Sin embargo, otras aportaciones como el análisis de fiabilidad mediante análisis de la varianza, la teoría de la generalizabilidad, o la orientación del espacio muestral de items, pueden ser incluidos actualmente como teoría 'clásica' de los tests por oposición a la otra gran familia de desarrollos de teoría de los tests que supone la teoría de la respuesta al ítem, pero en su día pudieron verse como modelos alternativos a la teoría clásica.

Quizás pueda reservarse la denominación de teoría clásica para esta familia amplia y diversa, pero bien avenida, de aproximaciones a la teoría de los tests, y usar otras expresiones más ajustadas, como 'modelo de los tests paralelos' (Yela, 1.984) o "teoría clásica de las puntuaciones verdaderas" o "modelo lineal de Spearman" (Santisteban, 1.989), para referirse al núcleo inicial de esta rama de la psicometría. La teoría clásica de tests (TCT) ha sido clasificada dentro de la teoría débil de las puntuaciones verdaderas, por oposición a la teoría fuerte de las puntuaciones verdaderas, y a la teoría de la respuesta al ítem (Lord y Novick, 1.968; Santisteban, 1.989).

FIGURA 12. ALGUNAS APORTACIONES PRINCIPALES EN EL DESARROLLO DE LA TEORIA DE LOS TESTS



En la *figura 12* se muestran algunas de las principales obras de la teoría de los tests. En la parte superior se señalan algunas de las principales aportaciones

a la Teoría Clásica, desarrolladas en su mayoría en la primera mitad del siglo. En la parte inferior se muestran algunas de las aportaciones más importantes a la Teoría de la respuesta al ítem, en su mayoría propias de la segunda mitad del siglo.

El modelo lineal de Spearman y el Test de Binet. En todo caso, aún concediendo una definición amplia de la llamada teoría clásica de los tests, no cabe duda que el principal componente genético e histórico de la misma, y de toda la teoría psicométrica de los tests en general, es el modelo lineal de Spearman. Puede citarse la fecha de 1.904 como punto de arranque histórico de este modelo. En ese año Spearman realizó una primera formulación que incluía los elementos básicos que lo caracterizaban como un modelo lineal aditivo, la teoría del error de medida, la fiabilidad y la fórmula de atenuación.

La teoría de Spearman nunca tomo de manos de este autor forma de un texto integrado y completo. En lugar de ello, la teoría fue formulada originalmente a través de un conjunto de trabajos fechados en 1.904, 1.907, 1.910 y 1.913, para destacar los que han sido considerados después como principales. Habría que esperar mucho tiempo, hasta las obras de Guilford y sobre todo de Gulliksen (1.950) para encontrar una exposición sistemática de la teoría de Spearman. Quizás la obra de Gulliksen, una de las piezas principales a todas luces de toda la historia de la psicometría, constituye la definición más clara y menos discutible de los contenidos de la teoría clásica de los tests, de medio siglo de teoría psicométrica. Un claro reflejo en castellano de los contenidos de la teoría clásica puede encontrarse en los 'Apuntes de Psicometría' del Profesor Mariano Yela, que han constituido sin duda uno de los principales instrumentos de formación psicométrica en nuestro país durante muchos años.

En 1.905, un año después del trabajo seminal de Spearman, se producía la primera edición del test de Binet y Simon para la medida de la inteligencia. Quizás, ni el test de Binet ni la teoría de Spearman hubieran llegado a

tener la relevancia que han adquirido históricamente de haberse dado lo uno sin lo otro. Debe resaltarse como la teoría psicométrica de Spearman, que carecía de una teoría psicológica definida a la que referirse, daba un soporte formal psicométrico a la medición de la aptitud intelectual, y como el test de Binet, que carecía de un soporte y una justificación metodológica y también de una teoría de la inteligencia bien elaborada, ofrecía la vía de aplicación lúcida y práctica de la teoría psicométrica de Spearman.

Pero desde luego, la importancia del modelo lineal de Spearman y de toda la teoría "clásica" de los tests no es meramente histórica. Como ya hemos resaltado, la teoría psicométrica clásica sigue constituyendo uno de los pilares fundamentales de la psicometría teórica y aplicada y sus procedimientos pueden seguir considerándose vigentes dentro de sus fines. El carácter acumulativo del conocimiento psicométrico y la vigencia de la teoría clásica queda bien patente cuando F. M. Lord (1.980) al referirse al desarrollo de la teoría de la respuesta al ítem, el otro gran tronco alternativo a la teoría clásica, dice expresamente que la nueva teoría en nada contradice a la antigua, aun cuando haga muy poco uso de las ecuaciones de aquella para desarrollar las suyas propias.

El núcleo de la teoría clásica de los tests es sin duda la teoría de la fiabilidad, alrededor de la cual se han estructurado las principales aportaciones de la misma. A la cuestión de la fiabilidad en la teoría clásica, vista con una perspectiva histórica, dedicaremos el próximo punto.

2. Una perspectiva histórica de la cuestión de la Fiabilidad y otros temas relacionados.

Orígenes del concepto. El término fiabilidad fue introducido por Spearman en un conjunto de trabajos que publicó en el American Journal of Psychology entre 1.904 y 1.913. Estos trabajos ("The proof and measurement of association between two things"; " 'General Intelligence' objectively determined and measured";

"Demostration of formulae for true measurement of correlation"; "Correlation calculated with faulty data" y "Correlations of sums and differences") constituyen su contribución principal a la teoría de los tests y ninguno de ellos fue enfocado de un modo explícito y exclusivo sobre la cuestión de la fiabilidad.

A partir de aquellos trabajos se desplegó una intensa investigación acerca los factores que afectan a la fiabilidad de un test y acerca de los métodos de formas paralelas, test-retest, y diversos métodos de partición del test que permiten calcular la fiabilidad. Los trabajos de Kelley (1.924), Muenzinger (1.927); Symonds (1.928); Anastasi (1.934); Adams (1.936); Kelley (1.942), Guttman (1.945); Cronbach (1.947) y Thorndike (1.947), entre otros muchos, avanzaron en esa linea durante la primera mitad del siglo.

Fiabilidad y longitud del test: La fórmula de Spearman-Brown. Partiendo del modelo de tests paralelos debía responderse la pregunta ¿qué nivel de fiabilidad podrá alcanzar un test por incremento de su longitud (aumentando el número de sus items) manteniendo las formas ampliadas rigurosamente paralelas a la inicial? Esta cuestión estaba en el interés de la teoría y era posible ensayar una respuesta precisa. El volumen 3 del British Journal of Psychology, publicado en 1.910, recogería dos trabajos que independientemente desarrollados llegaban a la misma solución. Charles Spearman, en un trabajo titulado "Correlation calculated with faulty data" llegaba en la página 290 a presentar las fórmulas para el caso de doble longitud y para el caso de longitud n veces la del test original. William Brown, en el artículo siguiente, titulado "Some experimental results in the correlation of mental abilities", en la página 299, llegaba a las mismas fórmulas. Desde entonces serían conocidas como las fórmulas o la fórmula de Spearman-Brown.

Discusión de la fórmula de Spearman-Brown. Inicialmente las fórmulas de Spearman-Brown no fueron muy bien acogidas, generándose un conjunto de trabajos que

ponían en cuestión la estimación de la fiabilidad que la formulas permitían. Holzinger publicó en 1.923 una "Note on the use of Spearman's prophecy formula for reliability", en el Journal of Educational Psychology, donde concluía que la fórmula para el caso general llegaba a sobreestimar de modo impreciso el valor de la verdadera fiabilidad. W.L. Crum el mismo año publicaba otra "nota" que también ponía en cuestión la fórmula de Spearman-Brown. Kelley en 1.924 y en 1.925 replicó a Crum, y, en general, un conjunto de trabajos en los años sucesivos, (por ejemplo, Wood en 1.926 o Gordon en 1.924,) vinieron a confirmar que las fórmulas iniciales eran aceptablemente precisas bajo los supuestos para los que habían sido desarrolladas. La cuestión estaba comúnmente en de qué modo las formas sucesivas seguían cumpliendo los supuestos de paralelidad. Slocombe, en una revisión de la evidencia realizada en 1.927, llegaba a la conclusión de que la cuestión de la paralelidad de los items estaba en el centro de la discusión. Durante estos años se elaboraron también una serie de tablas para facilitar el uso de la fórmula, como por ejemplo las de Edgerton y Toops de 1.928, o las de Cureton y Dunlap de 1.930.

Extensiones de la fórmula de Spearman-Brown. La fórmula de Spearman-Brown recibió posteriormente más usos e interpretaciones que fueron corroboradas por los datos. Por ejemplo, Gordon en 1.924 sugirió que la fórmula podría utilizarse para estimar el incremento en la fiabilidad que se produciría con el incremento de jueces. Muy posteriormente se sugirió que también podría predecir el incremento en la fiabilidad debido al incremento del número de elecciones. La relación entre número de alternativas y fiabilidad fue expresada en la fórmula 30 de Carroll, en un trabajo de 1.945 publicado en Psychometrika y titulado "The effect of difficulty and chance success on correlations between items or between tests".

Longitud del test y validez. La cuestión de la fiabilidad-longitud podía extenderse al impacto sobre la validez y también al modo de ajustar las longitudes relativas de varios tests para maximizar la validez del compuesto, cuestión está última tratada por Horst en 1.948. Cuando se realiza el estudio del efecto de la longitud del test sobre la validez puede advertirse que la fórmula de corrección por atenuación, deducida por Spearman (1.904, 1.907, 1.910 y 1.913) puede ser considerada como un caso particular, como la correlación entre un test y un criterio con infinitos elementos y fiabilidad perfecta. La primera formulación de la fórmula de corrección por atenuación, dada por Spearman en 1.904, fue seguida por una clara crítica de Pearson ese mismo año. La cuestión, según la revisión que Thouless efectuó en 1.939 parece residir en el modo en que la fórmula es concebida y utilizada.

Homogeneidad y fiabilidad. Otra cuestión acerca de las variaciones del coeficiente de fiabilidad era el efecto de la homogeneidad-heterogeneidad del grupo (de la muestra sobre la que efectúan los cálculos) sobre el coeficiente de fiabilidad. Es decir, los cambios que podrían esperarse en la fiabilidad en función de cambios en la varianza del grupo. Esta cuestión fue resuelta por un conjunto de autores en torno al comienzo de la década de los veinte. El primer trabajo que presentó fórmulas para realizar esa estimación, bajo el supuesto de la equivalencia de los errores de medida, fue el de Kelley en 1.921. Holzinger ese mismo año realizó una dura crítica al supuesto sobre el que se basaban esas fórmulas. En realidad no siempre puede suponerse que el error de medida permanece constante a través de las variaciones en capacidad del grupo, y en la medida en que el supuesto no sea cierto las fórmulas de estimación no podrían aplicarse. Kelley desarrolló su trabajo original en otros posteriores de 1.923 y 1.927, y otros muchos autores en los años siguientes se ocuparon de la cuestión. Por ejemplo, Otis, que puede considerarse el verdadero padre del Alfa Army Test, dedicó atención al tema en 1.922 y Thurstone en 1.931. Por su parte Toops y Edgerton (1.927), Cureton y Dunlap (1.929), Rulon (1.930) y Gulliksen (1.950)

elaboraron ábacos y tablas para facilitar el cálculo de la estimación de la fiabilidad en función de cambios en la varianza del grupo. El supuesto de igualdad del error de medida a través de los cambios en la varianza del grupo solo podría sostenerse si el error de medida permaneciese invariante frente a cambios en la magnitud de las puntuaciones del tests. Ello exige poder considerar el error de medida como función de la magnitud de las puntuaciones del tests, un problema que resolvió Mollenkopf en los años 1.948 y 1.949. La conclusión de esos análisis pone de manifiesto que para una distribución perfectamente simétrica y de curtosis 3 el error de medida medio es igual a través de las puntuaciones; sin embargo, para distribuciones sesgadas o leptocúrticas o platicúrticas, se espera que el error de medida varíe con la magnitud de las puntuaciones.

Homogeneidad y validez. Por otro lado, si se habían establecido los efectos de variabilidad del grupo sobre el coeficiente de fiabilidad sería posible buscar su extensión, como en el caso de la cuestión de la longitud, sobre el coeficiente de validez del test. El efecto de la homogeneidad o heterogeneidad del grupo sobre la validez fue establecido primero por T.L. Kelley en 1.923 en un libro titulado "Statistical Methods", y fue retomado por una serie de autores como Guilford en 1.942, Crawford y Burnham en 1.946, Garrett en 1.947 y también Thorndike en 1.947. Sin embargo, parte de la cuestión había sido anticipada por Pearson en 1.903 al deducir entonces la fórmula para la estimación de la correlación conocida la varianza de ambos grupos para la variable sujeta a selección explícita. Pearson también había presentado en 1.903 una fórmula para la estimación de la correlación entre dos variables selectivas en lo que se ha denominado caso de tres variables de la selección univariada; esa linea fue a su vez proseguida por Thorndike en 1.947. Por último, también fue Pearson en 1.903 quién extendió esas formulas al caso de selección multivariada, linea en la que proseguirían Aitken en 1.934 y Burt en los años 1.943 y 1.944.

Contraste de paralelidad. El modelo que había formulado Spearman a principios de siglo y que se estaba desarrollando intensamente se fundaba en el concepto de tests paralelos, de modo que, desde el punto de vista práctico, aplicado, debía resolverse la cuestión de discernir empíricamente si dos tests dados, construidos para ser paralelos o no, cumplían efectivamente las condiciones para ser considerados tests paralelos. Wilks resolvió esta cuestión en 1.946 creando un estadístico que llamó L_{mvc} , y que varía entre cero y uno, que resulta adecuado para poner a prueba simultáneamente la hipótesis de que todas la medias, todas las varianzas y todas las correlaciones de tests supuestamente paralelos son iguales. Si medias, varianzas y covarianzas, son iguales tal estadístico L_{mvc} vale 1. Wilks generó un conjunto de tablas para contrastar este estadístico para el caso de muestras pequeñas y señaló que el estadístico -N log e L_{mvc} distribuye como chi-cuadrado con (k/2)(k+3)-3 grados de libertad, para su contraste en muestras grandes. Desde este estadístico se desprenden las fórmulas necesarias para los casos en que se requiera la puesta a prueba de que, separadamente, medias, por un lado, o varianzas y covarianzas, por otro, son iguales.

Fiabilidad test-retest. Una cuestión importante era establecer bajo que condiciones una misma prueba se mantenía paralela a si misma a través del tiempo, es decir, enfrentar la cuestión de la fiabilidad como estabilidad temporal de los resultados de los tests. Como Woodrow puso de manifiesto en 1.932 los sujetos presentaban una variabilidad en su respuesta a los tests. Esta variabilidad afectaría a la correlación entre sucesivas administraciones de un test como una función de los cambios ocurridos a los sujetos a través del tiempo y no solo como función de la estabilidad del test mismo. En síntesis la cuestión podía plantearse en términos de cómo distinguir los cambios en las puntuaciones debidos a cambios reales en la variable a medir y los cambios en las puntuaciones debidos a inestabilidad del instrumento de medida. Obviamente ambos tipos de variabilidad se

presentaban en la práctica indisolublemente entremezcladas. Paulsen, en 1.931 fue el primero en proponer métodos para distinguir variaciones debidas a la inestabilidad del test de las debidas a la inestabilidad del rasgo. Su trabajo fue seguido por los de Thouless en 1.936 y 1.939, el de Preston en 1.940 y el de Jackson y Ferguson en 1.941. Por su parte Woodrow en 1.932 y Greene en 1.943 desarrollaron trabajos acerca de la estabilidad con retests usando la misma forma.

Fiabilidad dos mitades y fiabilidad pares-impares. El modelo de los tests paralelos se extendió también al caso en que una sola forma del test estaba disponible, para este caso se ingenió el procedimiento de dividir el test en dos mitades calcular la correlación entre ellas y aplicar la fórmula de Spearman-Brown para el caso de longitud doble. Una de las variedades, pero no la única, en que puede calcularse la fiabilidad mediante dos partes del mismo test es la conocida como pares-impares. Flanagan en 1.937, y Rulon en un trabajo de 1.939 titulado "A simplified procedure for determining the reliability of a test by split-halves", desarrollaron dos fórmulas, equivalentes a su vez entre sí, útiles para esta aproximación por partes para el cálculo de la fiabilidad. Por su parte Guttman mostró en 1.945 que la fórmula de Flanagan, (la lambda 4 de Guttman) era un límite inferior, y que equivalía a la fórmula de Spearman-Brown para el caso de longitud doble cuando se cumplía el supuesto de igualdad de varianzas entre las partes del test. De acuerdo con el trabajo de Guttman, las fórmulas de Flanagan y de Rulon tenderían a infraestimar la fiabilidad del test. Cureton en 1.931, y después Dunlap en 1.933 y Stephenson en 1.934, sugirieron partir el tests en más de dos partes para estimar la fiabilidad. Estas sugerencias aproximan el modelo split-half al análisis de la fiabilidad de los elementos.

Comparación de métodos. Durante los años treinta y primeros de los cuarenta se realizaron toda una serie de trabajos de comparación entre las diversas aproximaciones (pares-impares, formas paralelas, test-retest de la misma forma, test-retest de diferentes

formas y otras combinaciones). Entre esos trabajos pueden citarse el de Foran en 1.931, el de Jordan en 1.935, el de Goodenough en 1.936, el de Remmers y Whisler de 1.938, el de Ferguson de 1.941, el de Jackson y Ferguson del mismo año y el de Greene de 1.943. Todos estos trabajos dieron como resultado la certeza de que diferentes métodos de estimación de la fiabilidad arrojan resultados diferentes, siendo, en general, la fiabilidad por formas paralelas la más baja y la de pares-impares, con la corrección oportuna respecto de la longitud, la más alta.

Fiabilidad como homogeneidad o consistencia interna: la fórmula KR20 de Kuder y Richardson. Una aproximación distinta al cálculo de la fiabilidad de un test fue seguida por quienes estudiaron el problema desde el ángulo de la homogeneidad de los items. Desde la aportación de Johnson y Newman en 1.936, un conjunto numeroso de trabajos se ocuparon en utilizar el análisis de la varianza como un procedimiento de estudio de la fiabilidad. Entre ellos pueden mencionarse primero el de Hoyt de 1.941, y luego los de Jackson de 1.939, 1.940 y 1.942, el de Alexander de 1.947, y el de Jackson y Ferguson de 1.941. Esta aproximación, así como la seguida por Kuder y Richardson en su trabajo de 1.937, y Richardson y Kuder de 1.939, y también otros, se caracteriza por evaluar la fiabilidad de un test desde el punto de vista genéricamente denominado de la homogeneidad, o de la consistencia interna, que no utiliza el método de formas paralelas. Guttman (1.945; 1.946) consideró la fiabilidad buscando estimaciones de los límites inferiores de la fiabilidad de los tests. Por diversos caminos de análisis, y bajo diversos supuestos, un amplio grupo de autores llegó separadamente a diversas formulaciones, matemáticamente equivalentes, de una expresión de la consistencia interna que habitualmente se reconoce bajo la denominación de formula 20 de Kuder-Richardson publicada en Psychometrika en 1.937, pero que puede encontrarse también como la fórmula 11 del capítulo 16 de Gulliksen (1.950), la fórmula 29 del capítulo 5 de Jackson y Ferguson (1.941) y la fórmula lambda 3 de Guttman (1.945). Paradójicamente los supuestos utilizados para la obtención de esta familia de

reexpresiones de una fórmula diferían notoriamente entre sí. Por ejemplo, Harold Gulliksen en su "Theory of Mental Tests" de 1.950 desarrolla la forma general de la fórmula (ecuación 10), y la forma restringida para items dicotómicamente puntuados (ecuación 11), bajo el único supuesto de que la covarianza promedio entre los items no paralelos es igual a la covarianza promedio entre los items paralelos.

La fórmula KR21 y su relación con KR20. Kuder y Richardson en el mismo trabajo de 1.937 dedujeron la llamada KR-21, una aplicación de la KR-20 bajo el supuesto de igual varianza de los items dicotómicamente puntuados, o si se prefiere, para decirlo al modo usual, bajo el supuesto de igual dificultad de los items. También esta fórmula se halla desarrollada en parte del conjunto de trabajos antes citado. En el modo en que esta fórmula se trasmite actualmente desafortunadamente se ha perdido también su interpretación como un límite inferior de la fiabilidad. En efecto, conocidas solamente la media del test, la desviación típica de los items y el número de items (para un test con items dicotómicamente puntuados cuya puntuación total es igual a la suma de aciertos), la KR-21 significa el límite inferior de la fiabilidad. Si todos los items presentan la misma dificultad entonces KR-21 dará efectivamente el mismo resultado que KR-20, pero si los items difieren en dificultad entonces KR-20 será mayor que KR-21 (Gulliksen, 1.950).

Coeficiente alfa y su relación con KR20 y KR21. Como es sabido, este grupo de fórmulas constituyen solo un caso particular de la fórmula que nos hemos acostumbrado a denominar coeficiente alpha, que, a su vez, es fruto de la desigualdad de Cauchy-Schwartz aplicada sobre las tautologías que ponen en relación la varianza del test total con las de los items que lo forman aditivamente, y, que, como ya mostró primeramente Guttman, y expresa con toda claridad Lord (1.980) "no es un coeficiente de fiabilidad; es un límite inferior".

Dado que KR-20 equivale a alfa para items dicotómicamente valorados, y dado que alfa es menor o igual que el coeficiente de fiabilidad definido como proporción de varianza verdadera, entonces KR-21 es una expresión menor o igual que el coeficiente de fiabilidad, de modo que KR-21 que expresa un límite inferior de la misma por debajo del límite inferior que suponen alfa o KR-20.

Como puede advertirse en la discusión anterior la fórmula del coefiente alpha, y también esta concepción de la misma, estaba implícita o explícita en los diversos desarrollos que se habían producido para dar lugar a las diversas formulaciones de la KR-20. Estos trabajos son anteriores temporalmente a los de J.L. Cronbach de 1.947 y 1.949 acerca de la fiabilidad de los tests, y a su conocido trabajo de 1.951 "Coefficient alpha and the internal structure of tests", de modo que también esta fórmula puede considerarse un producto del trabajo de diversos autores entre los años 1.937 (fecha del trabajo de Kuder y Richardson) y 1.947 (fecha del trabajo de Cronbach). Por otra parte, es posible mejorar el resultado de la fórmula evitando algunos supuestos. Si el investigador conoce las correlaciones entre los items y el test total o las correlaciones o covarianzas entre los items, entonces no es necesario desarrollar la estimación de la fórmula bajo supuestos o restricciones acerca de estas correlaciones, y, consecuentemente, puede realizarse una estimación más aproximada de la fiabilidad. Kuder y Richardson en su trabajo de 1.937, y después Guttman en 1.945, se ocuparon de la estimación de la fiabilidad del test conocidas esas relaciones. No obstante, comúnmente se ha considerado la KR-20 suficientemente precisa, evitándose cálculos adicionales que actualmente, sin embargo, habrían dejado de ser costosos.

Análisis de los elementos. El desarrollo de los conceptos de fiabilidad no podía permanecer separado de la cuestión del análisis de los elementos de los tests. Para Gulliksen (1.950) "básicamente, el análisis de items se refiere al problema de seleccionar items para un test de modo tal que el test resultante tenga ciertas

características especificadas." Un trabajo pionero en este campo fue el de Lentz, Hirshstein y Finch de 1.932 sobre valoración de métodos de evaluación de items. Durante el primer tercio de siglo se desarrollaron una gran cantidad de índices para el análisis de items, sin embargo, los trabajos de revisión ponían de manifiesto una débil o inexistente formulación de la relación entre el índice referido al ítem y los valores de la fiabilidad y/o la validez del test a los que teóricamente debían servir. Guilford llega a listar y dar cuenta de noventa índices en su "Psychometric Methods" de 1.936, riqueza de la que deben descontarse numerosos índices redundantes o no bien fundamentados. Adkins en un trabajo de 1.938 clasificó los índices de análisis de items en tres grandes grupos: (a) los referidos a la correlación entre el ítem y el test, (b) los referidos a la inclinación de la regresión del test sobre el ítem, y (c), los referidos a la inclinación de la regresión del ítem sobre el test. No obstante, algunos trabajos particulares en este campo como el de Richardson de 1.936 sobre la relación de la dificultad con la validez diferencial, el de Horst de 1.934 sobre el método de residuales sucesivos para el análisis de items, o el de Adkins y Toops "Simplified formulas for ítem selection and construction", de 1.937, merecen mención especial.

Fórmula de corrección de la respuesta al azar. También debe mencionarse la aportación de Guilford al campo de la puntuación y corrección de tests al publicar en 1.936 en el trabajo "The determination of ítem difficulty when chance success is a factor", la fórmula de corrección de la respuesta al azar para pruebas con respuesta verdadera. Una fórmula que fue desarrollada bajo los supuestos de que todos los sujetos que no conocen la respuesta verdadera contestan al azar y de que al jugar al azar un sujeto que no conoce la respuesta tiene igual probabilidad de escoger cualquier alternativa. La dificultad principal con esta fórmula reside en que el supuesto de igual frecuencia para los distractores no se cumple habitualmente.

Antes de que Guilford presentara su fórmula, Horst, en un trabajo titulado "The difficulty of a multiple choice test item" publicado en 1.933, había presentado otra fórmula que permite estimar el porcentaje de personas que conocen la respuesta correcta a un ítem como razón entre el número de sujetos que eligen la respuesta correcta menos los que eligen el distractor más elegido, y, en el denominador, el número total de personas que responden al ítem. También se desarrollaron un buen número de métodos orientados al cálculo de la dificultad seleccionando un porcentaje (entre el 10% y el 33%) inferior y superior de la muestra. En todo caso estas estimaciones pueden darse por innecesarias bajo los actuales procesos de cálculo, pero algunas como las de Flanagan de 1.939 y las de Davis de 1.946, han tenido un papel importante en el trabajo práctico (junto con los ábacos, las tablas y los métodos abreviados) hasta hace unos pocos años.

Relación de la dificultad del ítem con su fiabilidad y validez. Otros trabajos se orientaron a estimar que nivel de dificultad de los items de un test favorecía su fiabilidad y validez. El resultado más notorio de esos trabajos, que puede encontrarse en los de Cook de 1.932 y T.G. Thurstone del mismo año, consiste en que los tests con items cuyo índice de dificultad está en torno a 0'50 suelen presentar la validez más elevada. Carroll, por ejemplo, estudió este efecto en un trabajo de 1.945. Por otro lado, las correlaciones simples, parciales y múltiples entre items y criterio también fueron durante mucho tiempo demasiado costosas, incluso para los viejos ordenadores, lo que llevó al desarrollo intensivo de toda una literatura para la economía de tiempo y operaciones de cálculo (a costa habitualmente de estimaciones, supuestos, y restricciones solo justificables por razones de economía) que actualmente no merece la pena considerar.

Puntuaciones estandarizadas. A principio de la década de los sesenta se desarrollaron dos simposiums relacionados con la teoría de tests, publicándose los trabajos de ambos en Educational and Psychological Measurement. El primero se ocupó del uso de las

puntuaciones estandarizadas para tests de aptitudes y logro. El segundo se refirió a la cuestión de los límites de tiempo y sus efectos sobre las puntuaciones de los tests. En el primero, se pusieron de manifiesto algunas discrepancias respecto a la cuestión de la estandarización de las puntuaciones. Así, mientras que Wesman (1.962), Gardner (1.962), Ebel (1.962) y Flanagan (1.962) defendieron de uno u otro modo la práctica de utilizar puntuaciones estandarizadas, Angoff (1.962) representó el punto de vista opuesto, defendiendo puntuaciones con origenes arbitrarios y unidades sin significado. A su juicio, el intento de dotar a los tests de unidades y origen con un significado favorece una mala interpretación de las puntuaciones de los tests.

Construcción de tests. Una exposición sistemática y ordenada de todas estas cuestiones desde el punto de vista metodológico y teórico, más que histórico, puede encontrarse en la obra de Harold Gulliksen "Theory of Mental Tests". Gulliksen (1.950) presentó una síntesis de las, a su juicio, cinco principales cuestiones que deben enfrentarse desde el punto de vista de la construcción de los tests mentales:

1. Redactar y seleccionar los items de los tests.

2. Asignar puntuaciones a cada sujeto que responde al test.

3. Determinar la precisión o fiabilidad de las puntuaciones que arroja el tests.

4. Determinar el valor predictivo o validez de las puntuaciones del test.

5. Comparar los resultados con aquellos obtenidos usando otros tests o otros grupos de sujetos. En estas comparaciones deben tenerse en cuenta los factores de longitud del tests y grado de homogeneidad o heterogeneidad de la muestra.

Este resumen de problemas es a la vez un programa de trabajo y una apretada síntesis de la teoría clásica de tests, un índice conceptual de la misma obra de Gulliksen de 1.950, que es uno de los pilares de la psicometría clásica, y, a la vez también, un primer algoritmo de pasos necesarios en la elaboración de una prueba.

3. La cuestión de la validez.

La cuestión de la validez siempre ha aparecido algo más difusa y menos inequívocamente establecida dentro de la aproximación clásica a los tests. Las dificultades para ordenar estructuradamente los diversos conceptos y tipos a que la validez se refiere han dado lugar a diferentes clasificaciones y sobre todo un buen número de matices y divergencias en aspectos puntuales. Sin embargo el núcleo conceptual básico de los tipos principales de validez puede considerarse bien establecido.

Durante la década de los cuarenta la insatisfacción con la definición de validez y la aplicación de la misma estaba presente en el espíritu de un buen número de trabajos. J.G. Jenkis titulaba explícitamente un artículo publicado en 1.946 "Validity for what?" Por una parte, se había postulado que la validez era un criterio de bondad principal que debía apoyarse en una justificación empírica explícita, por otra, un buen número de conceptos psicológicos no se prestaban fácilmente a identificar los criterios externos definidos sobre los que fundar la validez. Además, el cuerpo teórico acerca de la validez podía considerarse, comparado con el articulado y complejo aparato desarrollado acerca de la fiabilidad, del todo insuficiente.

En este clima un conjunto de autores se ocuparon de la validez generando un buen número de nuevos conceptos. En 1.947 Mosier revisaba el concepto de validez aparente y en 1.951 se ocupaba de la cuestión de la validación cruzada. Gulliksen acuñó el término de "intrinsic

validity", en un artículo de 1.950. El mismo año Florence L. Goodenough introducía la importante distinción entre tests tomados como signos y tests tomados como muestras. Y también en 1.950 Anne Anastasi dedicaba el artículo "The concept of Validity in the interpretation of tests scores" a discutir el significado de la validez y su papel entre los criterios de bondad de los tests. Cronbach en su libro de 1.949 "Essentials of Psychological Testing" introducía la distinción entre la validez lógica y la validez empíricamente comprobada. Y Guilford, en un trabajo de 1.946 que pretendía esclarecer los procedimientos para el establecimiento de la bondad de un test, aportaba como novedad el concepto de validez factorial. Un cúmulo de conceptos que nacían ahora o que retomaban aspectos de investigaciones muy anteriores pero que, en cualquier caso, estaban en ebullición mostrando una intensa preocupación por delimitar el concepto de validez, sus tipos y los procedimientos de obtención admisibles.

Ante esta situación, entre 1.950 y 1.954 un comité para los tests psicológicos constituido por la APA y formado por Bordin, Challman, Conrad, Cronbach, Humphreys, y Meehl, estuvo trabajando con la meta de determinar y unificar 'oficialmente' el conjunto de criterios que un test debe cumplir para su uso profesional. Las principales conclusiones de aquellos trabajos fueron publicadas en las "Technical Recommendations for Psychological Tests and Diagnostic Techniques", que aparecieron como un suplemento del Psychological Bulletin en 1.954. Retomando y ordenando los conceptos anteriores se llegaron a establecer cuatro grandes tipos de validez: la validez predictiva, la concurrente, la de contenido y la de constructo. La validez predictiva y la concurrente eran dos formas de la validez criterial, distintas en función de la distancia temporal entre la medición del test y la del criterio. La validez de contenido hacia referencia al grado en que los elementos de un test pueden demostrarse como una muestra representativa del contenido de un universo bien delimitado. Por último, la validez de constructo implicaba un concepto más ambicioso que permitía extender el concepto de validez y, bajo alguna interpretación, reunir

los otros tipos de validez bajo una misma lógica. En realidad el concepto de validez de constructo había nacido antes en un subcomité de la APA formado por P. E. Meehl y R. C. Challman, y sería especialmente interpretado, desarrollado y divulgado a partir de un importante artículo de Cronbach y Meehl "Construct validity in Psychological Test" que apareció en el Psychological Bulletin en 1.955. Para sintetizar la lógica generadora de la validez de constructo pueden citarse las palabras de estos autores "El proceso de investigación de la validez de constructo no difiere esencialmente de los procedimientos generales usados en ciencias para establecer y confirmar teorías" (Cronbach y Meehl, 1.955). Este punto de vista implica considerar a los instrumentos de medida como hipotéticas mediciones que deben ponerse a prueba como cualquiera otra hipótesis. En medio, antes de esta síntesis, se había discutido el concepto de red nomológica de constructos y los diversos procedimientos concretos en que esta puesta a prueba de los tests podía llevarse a cabo.

El concepto de validez de constructo se convirtió en una herramienta intelectual para entender, juzgar y ordenar otras concepciones de la validez. Por ejemplo, toda la temática de la dimensionalidad de las medidas podía recogerse ahora como un aspecto de la validez de constructo relativo a la estructura, generalmente determinada por análisis factorial o por análisis de componentes principales, del test. La incorporación del análisis factorial en sentido amplio dentro del marco de la validez implicaría que los sucesivos desarrollos de esta técnica tendrían una importancia psicométrica explícita. Quizás los manuales clásicos más extendidos de análisis factorial sean el de Harman de 1.967 y el de Horst de 1.965, recomendables solo después de otros suficientemente serios pero más introductorios como el de Comrey (1.985) o el de Kim y Muller (1.978).

Paralelamente al desarrollo de los métodos factoriales se fueron desarrollando a partir de trabajos pioneros de Wright (1.934) un conjunto de procedimientos que, aunque ha recibido también otros nombres, puede

conocerse bajo la etiqueta de path análisis. Los modelos factoriales, como modelos de dimensionalidad latente de las medidas, y los modelos path, como modelos de la estructura causal de las relaciones entre variables, fueron integrados, principalmente por Jöreskog, en los llamados modelos de estructuras de covarianza, a veces denominados también como modelos LISREL, tomando las siglas que se usan para la denominación del programa. Pues bien, los modelos de estructuras de covarianza o modelos LISREL, en tanto que representan un análisis de la dimensionalidad y la estructura de relaciones de causalidad en que las mediciones están ubicadas, se han incorporado en los últimos años por derecho propio como procedimientos de aproximación a la validez de constructo de las mediciones psicológicas.

4. Generalizabilidad.

La teoría clásica de los tests, con su clara distinción entre fiabilidad y validez como aspectos separados, y su capacidad para arrojar distintos valores para el coeficiente de fiabilidad y para el coeficiente de validez sin una teoría integradora que explique todas las variantes posibles en los resultados, fue acumulando un número de críticas que pueden encontrarse reflejadas en parte en el trabajo de Tryon (1.957). Sin embargo quizás ha sido el poderoso núcleo de investigadores en torno a la figura de Cronbach el que más atención prestó a esta cuestión durante la década de los sesenta, llegando a retomar aquellas críticas y formular una nueva visión del campo de la fiabilidad que ha sido conocida bajo el neologismo de "theory of generalizability". Recientemente Santisteban (1.989) disponiendo de información directa del mismo Cronbach, ha explicado el significado exacto y el origen de este neologismo.

El cuerpo teórico principal de la teoría de la generalizabilidad quedó establecido en la primera mitad de los sesenta en "Theory of generalizability: A liberalization of reliability theory" de Cronbach, Rajaratnam y Gleser

(1.963), "The signal-noise ratio in the comparison of reliability coefficients" de Cronbach y Gleser (1.964), y en "Generalizability of Scores influenced by multiple sources of variance" de Gleser, Cronbach, y Rajaratnam (1.965), si bien tuvo una de sus expresiones más acabadas en un volumen de Cronbach, Gleser, Nanda y Rajaratnam titulado "The dependability of behavioral measurements: theory of generalizability". Sucintamente, la teoría de la generalizabilidad implica la concepción de cada medición como un punto en un espacio definido por un número de facetas o universos de generalización que pueden presentarse bajo diversas condiciones, de modo que una aplicación del análisis de varianza permite determinar los componentes de la varianza atribuibles a las diversas condiciones. Los estudios que se orientan a establecer los componentes de varianza se denominan estudios G o de generalización, y a partir de ellos se obtienen coeficientes G definidos como una razón entre la varianza de las puntuaciones del universo y la varianza de las puntuaciones observadas. Tales coeficientes G son los equivalentes de los coeficientes de fiabilidad baja la teoría de la generalizabilidad (Martínez-Arias, 1.981; 1.984).

Históricamente, la aplicación del análisis de varianza al estudio de la fiabilidad no constituye una originalidad de la teoría de la generalizabilidad. El trabajo pionero en este campo es el de Hoyt de 1.941, que fue retomado por Gulliksen en 1.950 y seguido por Lindquist en 1.953, y sobre todo por Burt en un trabajo de 1.955 titulado "Test reliability estimated by analysis of variance". Evidentemente la teoría de la generalizabilidad no es una mera extensión de los trabajos de Hoyt, pero pone de manifiesto la capacidad de Cronbach para retomar, más que inventar, conceptos y teoría anteriores, y dotarlos de un nuevo lugar o un nuevo significado, como sucedió también con la fórmula alfa, extensión general de los trabajos de un buen número de autores anteriores.

Desafortunadamente, como se ha puesto de relieve reiteradamente, la formulación de Cronbach y colaboradores de la teoría de la fiabilidad contenida en los

trabajos citados, especialmente en el de 1.972, no estaba orientada hacia el cálculo empírico y aplicado de los coeficientes. Recientemente, Tourneau y Cardiner (1.985) han ofrecido una aproximación a la teoría de la generalizabilidad más orientada al cálculo práctico de sus coeficientes. La cuestión ya había sido resuelta satisfactoriamente en el capítulo dos, escrito por Cardinet y Allal "Estimation of Generalizability parameters", del libro de Leslie J. Fyans de 1.983. La obra de Fyans "Generalizability Theory: Inferences and Applications" es uno de los tratamientos más extensivos y completos de la teoría. Con anterioridad Shavelson y Webb habían realizado en 1.981 una revisión sobre la teoría de la generalizabilidad desde la publicación del libro de Cronbach y colaboradores de 1.972 hasta su propio trabajo. Esos trabajos han contribuido a extender la aplicación de la teoría de la generalizabilidad y a fomentar un número de desarrollos concretos. A principios de la década de los ochenta los trabajos de Brennan y colaboradores (1.980) y de Cardinet y colaboradores (1.981) contribuyeron por su parte a formular extensiones más complejas de la teoría. Recientemente, Allal (1.988) ha realizado una exposición sumaria que incluye una revisión de los principales avances de la teoría de la generalizabilidad.

Una de las repercusiones más interesantes del concepto de generalizabilidad ha sido su impacto sobre la interpretación de las relaciones entre la fiabilidad y la validez. El modo en que la relación existente entre fiabilidad y validez era considerada en la teoría de los tests se centraba en la restricción que la fiabilidad de los tests y las variaciones en esta podían presentar sobre la validez. Desde luego la teoría de los tests realizó desde el principio grandes progresos en esta linea formulando de un modo preciso esa relación, sin embargo, otros modos de estudiar esta relación son posibles. Fiske (1.971) ha puesto de manifiesto con toda claridad la forma clásica de enfrentar la cuestión y como la relación que puede establecerse entre fiabilidad y validez puede verse de otro modo: "En la bibliografía sobre la medición, fiabilidad y validez son tratadas usualmente en forma separada. Aunque se

reconoce generalmente que la fiabilidad es crucial porque pone un límite a la validez, la complementariedad fundamental de los dos tópicos se ha perdido en las discusiones. Ambos se refieren a aspectos de generalizabilidad. [...] La fiabilidad y la validez están realmente sobre un continuo de generalizabilidad. [...] La fiabilidad se refiere a la generalización del test como test; la validez se refiere a la generalización más allá del test hacia un criterio concreto o hacia un constructo." Campbell y Fiske (1.959) ya habían señalado que "la fiabilidad es el acuerdo entre dos esfuerzos para medir el mismo rasgo a través de métodos máximamente similares. La validez está representada en el acuerdo entre dos intentos por medir el mismo rasgo a través de métodos máximamente diferentes". Silva (1.982) ha representado las principales aproximaciones a la obtención de los criterios de bondad en la teoría clásica de los tests a modo de círculos concéntricos simbolizando la creciente pretensión de generalización. De dentro a fuera Silva ubica sucesivamente los métodos de consistencia interna (inclusive los de partición), la estabilidad temporal (test-retest), las formas paralelas, la validez criterial (convergente), la validez criterial (predictiva) y la validez conceptual o de constructo. Como hemos visto, la teoría de la generalizabilidad tiene su propio modo de definir y calcular sus coeficientes G, sin embargo la lógica de la generalización ha permitido reordenar las aproximaciones clásicas bajo sugerencias como las recién mencionadas de Fiske o de Silva.