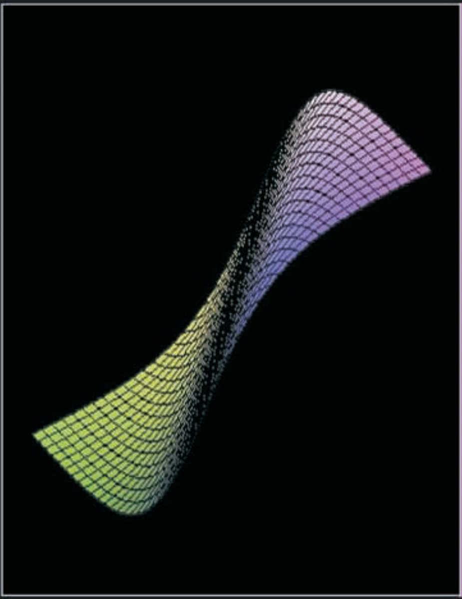
36196AA01A01



LA VALIDEZ EN LA MEDICIÓN PSICOLÓGICA

JESÚS M.ª ALVARADO IZQUIERDO
CARMEN SANTISTEBAN REQUENA

UNED
EDICIONES

LA VALIDEZ EN LA MEDICIÓN PSICOLÓGICA

AULA ABIERTA

Jesús Mª Alvarado y Carmen Santisteban

LA VALIDEZ EN LA MEDICIÓN PSICOLÓGICA

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

AULA ABIERTA (36196AA01A01)

LA VALIDEZ EN LA MEDICIÓN PSICOLÓGICA

Librería UNED: Bravo Murillo, 38; 28015 Madrid
Tels.: 91 398 75 60/73 73, e-mail: libreria@adm.uned.es

ISBN: 84-362-5061-3
Depósito legal: M-32.697-2006

Primera edición: noviembre de 2006

Impreso en España - Printed in Spain
Imprime: LERKO PRINT, S.A.
Paseo de la Castellana, 121. 28046 Madrid

ÍNDICE

PRÓLOGO

En esta monografía se aborda el problema de la validez en la medición de constructos psicológicos desde la perspectiva poco habitual de tratar conjuntamente la problemática de la medición científica y la de la validez de las medidas de los constructos psicológicos. Durante varias décadas se han ido desarrollando procedimientos para la validación de los instrumentos psicométricos que no prestaban atención a la cuestión primordial de si las medidas sobre las que se realizaban los análisis estadísticos habían sido o no obtenidas de modo "legítimo", puesto que lo relevante no era tanto si la medición reunía las características para científicamente considerarla como tal, sino más bien si las puntuaciones obtenidas eran o no útiles para propósitos concretos. Afortunadamente, este modo de proceder considerado por Michell (2000) como propio de una "ciencia patológica" parece llegar a su fin, gracias, entre otros muchos factores, al desarrollo y la sofisticación en los procedimientos estadísticos para la evaluación de la validez de los constructos. Obviamente, si la medida es débil, resultará extraordinariamente difícil mostrar evidencias sólidas de validez, con independencia del propósito al que iba dirigida esa medición. Por lo tanto, resulta mucho más coherente y económico dotarse de buenas medidas al iniciar el estudio, en lugar de confiarlo todo al azar o la intuición del constructor del test. En el marco de la Teoría de la Respuesta al Item se dispone actualmente de excelentes modelos que permiten obtener medidas de calidad (Borsboom y Mellenbergh, 2004) en las que la representación del constructo, aspecto clave para mostrar su validez, queda garantizada.

Esta obra comienza con una breve introducción, en la que se aborda de forma somera la cuestión de la medición en las denominadas ciencias físicas, incidiendo de forma más pormenorizada en la evolución y la problemática de la medición en Psicología. El último apartado del capítulo introductorio se dedica al problema de la validez desde una perspectiva histórica, haciendo mención explícita a las diferentes definiciones y normas que han ido apareciendo en su evolución.

El segundo capítulo se dedica a exponer las nociones más básicas de la teoría de la medida y de la medición en Psicología, comenzando por los conceptos métricos: métrica, espacios métricos y funciones medibles, para pasar posteriormente a los de estructuras, representación y escalas de medida.

En el tercer y último capítulo se estudian las propiedades métricas y estadísticas de algunos modelos psicométricos, en especial se refieren aquellos que han mostrado que poseen los requisitos que se exige a la medición científica, incidiendo en que esos modelos aportan una base sólida para la construcción de medidas válidas.

Con el fin de que sirva de ayuda en la comprensión de algunos apartados, se ha incluido un glosario de términos. Las limitaciones de esta obra hacen que no sea posible, ni quizás conveniente, que se incluyan los fundamentos de todos los términos y conceptos que se utilizan. Por ello, es probable que el lector no familiarizado con estos temas, tenga que acudir a otras fuentes. Creemos que esta tarea, cuando se trata de un alumno de estudios avanzados, de master o de doctorado, debe de formar parte de su proceso de formación.

INTRODUCCIÓN

1. La medición en Psicología

La mayor parte de los modelos matemáticos que se desarrollan en Psicología con anterioridad y durante los años 50 del pasado siglo, se adentran en el campo de los números reales cuando tratan de hacer ciencia, ya que la introducción del lenguaje cuantitativo que está relacionado con la medición, es lo que comúnmente se considera como garantía, e incluso como piedra angular del establecimiento del método científico. En sus publicaciones, los psicólogos proponen mediciones cuantitativas y en esa época se encuentra que prolifera el establecimiento de escalas psicológicas. Esto ocurre muy especialmente en Psicofísica, Psicometría y Aprendizaje, que se remontan a finales del siglo XIX y principios del XX con figuras tan destacadas como Weber (1795-1878), Fechner (1801-1887), Donders (1818-1870), Galton (1822-1911), Ebbinghaus (1850-1910) y Thurstone (1887-1955). En estos trabajos pioneros, lo que se tiene en mente es una escala real-valuada como las que se usan en Física para las medidas de ciertos atributos, y que han sido tan fructíferas en el desarrollo científico. Este intento, sin embargo, llevó en algunos campos de la Psicología a conclusiones erróneas y a confusiones teóricas. Todo ello conllevó a que se exploraran otros dominios numéricos que dieran cuenta de mejor manera de las necesidades y características especiales que pudiera tener la medición en Psicología y para el establecimiento de teorías psicológicas, formalmente establecidas, formuladas en términos matemáticos, pero no necesariamente en términos cuantitativos en un sentido tan restrictivo. Comienzan entonces a proliferar los estudios con una concepción más amplia del mundo de lo numérico en que se contemplan los números reales desde una perspectiva más general y desde un punto de vista más abstracto que la simple representación numérica de una medida o de un valor de escala. Es entonces cuando en Psicología se da el paso adelante en cuanto a que la distinción estricta entre lo cualitativo y lo cuantitativo pasa a ser un viejo principio filosófico, pues ahora en su ámbito de estudio el concepto de *cantidad* se ve a la luz de otras premisas y desde otras perspectivas donde lo cuantitativo, o el concepto de cantidad, se puede recalificar desde el punto de vista de la Teoría de la Medición Psicológica.

La consideración de la importancia que el desarrollo de una teoría de la Medición tiene para la Psicología, para que ésta mereciera en el futuro el nombre de científica, impulsó una gran cantidad de estudios sobre el tema, especialmente en la segunda mitad del pasado siglo. Consideramos que tiene su punto álgido en las décadas que van desde mitad de los 50 a finales de los 80, donde hay tal cantidad de investigaciones y publicaciones sobre este tema, con tan marcada presencia en todos los ámbitos y muy especialmente en las reuniones científicas que, por sobreestimación, casi podría inducir a considerar a la Teoría de la Medición Psicológica como sinónimo de Psicología Matemática.

Se puede considerar que el paradigma de medición que publican Scott y Suppes en 1958, explicando el paradigma representacional de Helmholtz (1887) de una forma general, precisa y estructurada, es un pilar básico de la actual concepción de la teoría matemática de la medición en Psicología.

El conocido paradigma representacional se basa en la noción de *representación* (esto es, que exista un isomorfismo u homomorfismo) de alguna estructura llamada empírica, o cualitativa, en alguna estructura numérica.

Las llamadas ciencias físicas no tienen dificultad en considerarse cuantitativas y el concepto de medición en ellas es tan fundamental y esta tan arraigado, que es difícil imaginar que pensaran prescindir de él. La Psicología y las Ciencias Sociales, sin embargo, en muchos casos lo han ignorado. Carnap (1966, p.52) dice que los conceptos comparativos son herramientas mucho más efectivas para describir, predecir y explicar que los clasificatorios, pues nos indican como un objeto se relaciona con otro en términos de "más que", "menos que". Sin embargo, es indudable que, siguiendo sus mismos ejemplos, decir que el objeto *a* está caliente, es mucho más informativo que decir que el objeto *a* está mas templado que el *b*, salvo que hayamos hecho asunciones o tengamos conocimiento acerca de cómo de caliente está el objeto *b*. Por supuesto, las relaciones comparativas nos pueden ayudar a describir, predecir y explicar, pero de una forma muy débil y para avanzar en el conocimiento, es necesario acudir a relaciones y estructuras más sólidas.

Cuando en Psicología y en general en las Ciencias Sociales y de la Conducta se ha intentado formalizar algún concepto o alguna teoría, se ha acudido a la observación y o a la experimentación. En la práctica, lo cuantitativo y/o medible se ha utilizado al menos en dos vertientes diferentes. Por un lado, en el mismo sentido que las medidas físicas, cuando se trata con medidas que son distancias, tiempos de reacción, intensidad de las señales, etc. Por otro lado, se han desarrollado procedimientos propios y peculiares que conllevan ciertos índices y/o escalas de medida, que no se pueden interpretar fuera de su contexto conceptual, como son, por ejemplo, los índices de sensibilidad individual al ruido, los de introversión/extraversión, el IQ de inteligencia, las escalas de aptitud o las de actitudes.

Existe y ha existido cierta controversia, muy marcada en algunas épocas, no solamente en relación con la naturaleza de la medición, sino también con el interés e intensidad con el que la comunidad científica aborda el tema, y concretamente la filosofía de la ciencia (Krantz, Luce, Suppes, y Tversky, 1971; Kyburg, 1984, Santisteban, 2003).

Krantz et al. (1971) dicen que hay una gran dispersión en la literatura, pues en Economía, Matemáticas, Filosofía, Física, Psicología y Estadística existen sistemas de axiomas y teoremas que intentan explicar cómo algunos de los atributos de los objetos, de la esencia de ciertas cosas y los sucesos se pueden representar numéricamente de forma razonable y esos resultados son los que constituyen los fundamentos matemáticos de la medición. Hace más de una veintena de años, Kyburg (1984) opinaba que, dada la prevalencia de la medición en todas las ramas de la ciencia, podríamos esperar que las discusiones sobre medición jugaran un importante papel en los escritos sobre filosofía de la ciencia. Esta expectativa la considera frustrada. En un debate actual sobre la influencia del método y/o los métodos en el desarrollo de la ciencia (Santisteban, 2003) también se pone de manifiesto la aparente disociación entre el desarrollo científico y el de la filosofía de la ciencia al tratar estos temas, teniendo escasa presencia en la formación de los científicos. En la presente obra se hacen puntualizaciones acerca de la cuantificación, la medición, la búsqueda y establecimiento de estructuras y la formalización.

En relación con ese aparente desinterés, Kyburg dice que puede significar una de las dos cosas siguientes: que la medición se ha llegado a comprender tan bien y es tan fácil de entender que no se necesita decir mucho más acerca de ella; o que los problemas de la medición no son ni siquiera suficientemente apreciados. Es con ésta última razón con la que concuerdan Ellis (1968) y Koslow (1982). Ellis dice que tiene que remontarse a los años 20 para encontrar grandes trabajos que traten la medición y que sean de naturaleza filosófica, refiriéndose concretamente a los trabajos de Campbell (1920, 1928) y a los de Bridgman (1922). Compartiendo esta opinión Koslow, sin embargo, dice que la literatura sobre el tema está en auge entre sus contemporáneos y que en esos años se había escrito mas sobre medición que antes de mediados de los 60. Entre los grandes autores clásicos sobre medición cita a Mach (1960) y a Helmholtz (1977).

Autores como Krantz, Luce, Suppes, Tversky, Zinnes, Pfanzagl y Nagel, entre otros, comparten el otro punto de vista, implícitamente al menos, el de aquellos que opinan que la medición es un tema perfectamente comprendido. Las obras de estos autores hablan por si solas de que la medición es un tema bien entendido en el que, sin embargo, hay muchas cuestiones de fondo complejas y muchas cuestiones técnicas, algunas de ellas de gran dificultad, que hay que explorar e investigar y a ello dedican obras muy relevantes, tanto por su extensión como por su profundidad.

No obstante, en la literatura hay una ingente cantidad de trabajos, que generalmente aparecen en revistas de Psicología, que suelen centrarse en problemas técnicos muy concretos refiriéndose a como se debe hacer la medición de un cierto atributo, o construyendo algún tipo de escala, pero sin llegar a establecer los fundamentos matemáticos que den validez a la medida de ese atributo. Es decir, se quedan en la parte técnica del problema concreto, sin llegar a establecer los pertinentes axiomas que permitan la generalización de los resultados. Nos referimos concretamente al establecimiento de los axiomas que abocarán a que se cumplieran al menos dos teoremas: el *teorema de representación*, que muestra que el atributo se puede representar mediante una cierta estructura de los números reales, y el *teorema de unicidad*, que muestra como se relacionan cualesquiera dos funciones (de los objetos en los números reales que representan el atributo en cuestión), por ejemplo, si entre ellos existe una relación lineal, o una transformación monótona, etc. Por ejemplo, el atributo "peso" de los objetos se representa en el campo de los números reales mediante una escala de razón, pasándose de unas unidades a otras multiplicando (o dividiendo) por una constante y adicionalmente, los pesos de pares de objetos, cualesquiera que sean, se combinan de forma aditiva, siendo esa suma también un número real.

El problema del significado debe de estar presente en todo el proceso, validando todos y cada uno de los pasos en la investigación.

2. Psicometría y medición

La Psicometría, desde un punto de vista conceptual y muchas veces incluso desde su punto de vista instrumental, tiene uno de sus pilares en los conceptos más rigurosos de la medición, y toda su estructura es un tipo de medida psicológica. De hecho, algunos libros de Psicometría, como el de Allen y Yen (1979) lleva por título "Introduction to Measurement Theory" y el titulo del libro de Wainer y Messick (1983) en homenaje al psicómetra Lord es "Principals of Modern Psychological Measurement". En el libro de Allen y Yen, las autoras intentan que esa obra sirva de texto y libro de referencia para aquellos que usen o construyen tests psicológicos e interpreten las puntuaciones de los tests y las escalas. Consideran que los libros sobre medición tienden a ser de tres tipos: (1) volúmenes que son matemáticamente rigurosos y extremadamente difíciles, (2) libros que discuten las propiedades estadísticas de las mediciones de una forma poco rigurosa, en forma de recetas, como un libro de cocina, y (3) libros que discuten los diferentes tipos de tests y la filosofía de la medición. En esa obra, siguiendo a Stevens (1951), se dice que medición es la asignación de números a los individuos de una forma sistemática, como un medio de representación de las características de los individuos. Se considera la Teoría de la Medida como una rama de la Estadística Aplicada que intenta describir, categorizar y evaluar la calidad de las medidas, mejorar su utilidad, exactitud, su significado, así como proponer métodos para que se desarrollen nuevos y mejores instrumentos de medida.

En el prólogo de una obra muy específica de Psicometría, especialmente dedicada a un tipo especial de teoría de tests y a sus aplicaciones "Applications of Item Response Theory to Practical Testing Problems", su autor, Frederic M. Lord, dice:"*the purpose of this book is to make it possible for measurement specialists to solve practical testing problems...*". Al inicio del apartado anterior, hemos citado quienes son los verdaderos precursores de la Psicología como ciencia positiva, y la eclosión de la teoría de la medición, entendida como disciplina. Hay autores que la sitúan en los años 30 del pasado siglo, que es cuando se funda la revista Psychometrika (1935) y muy poco después Educational and Psychological Measurement (1941) y el British Journal of Statistical Psychology (1947). Por lo tanto, creemos que cuando se invoca la Teoría de la Medición, hay que tener presente a la Psicometría y a su contribución. Esto no conduce a considerar que todo lo que conlleva e implica la Teoría de la Medición esté contemplado y contenido en la Psicometría, sino que bebe de sus fuentes. Las bases rigurosas de la Medición deben de tener presencia en todos los ámbitos de la Psicología en los que se quieran desarrollar teorías formalmente construidas y con modelos experimentalmente contrastables.

En muchos contextos psicológicos, y muy especialmente en el de la Psicometría, se han propuesto modelos de medición que han sido de gran transcendencia y que han supuesto un despegue teórico e instrumental en el contexto en el que se han generado. Al respecto, Susan Embretson asegura que, en nuestros días, la teoría de la respuesta al ítem (TRI) es, con diferencia, una de las tecnologías más relevantes de medición aplicada (Embretson, 1996; Daniel, 1999). No cabe duda de que esto sea cierto y que se debe fundamentalmente a las características del modelo de medición generado por Rasch.

Rasch (1960) descubre que la única forma de estimar la aptitud lectora de los sujetos, comparando el rendimiento de los sujetos en diferentes tests de lectura oral, es aplicar las propiedades de aditividad de las distribuciones exponenciales. Rasch usa como modelo las distribuciones probabilísticas de Poisson, que datan de 1837, por sus propiedades de aditividad e independencia.

El gran éxito de la formulación de Rasch consiste en que propuso un modelo matemático-probabilístico en el que la solución pasaba porque el modelo de medición debía restringirse sólo a formulaciones que produjesen estadísticos suficientes para los parámetros. El objetivo último es la estimación del parámetro "aptitud lectora", que es una variable latente, que requiere una medición indirecta. Rasch hace esa medición indirecta a través de una medida directa, que es el "número de errores" de un sujeto j, en dos tests o textos diferentes i, k. Considera que x_{ij}, el número de errores del sujeto j en el texto i, es función de su aptitud θ_j y que también lo es el número x_{kj} de errores de ese sujeto en el texto k. Usando el número total de errores, $x_{.j}$, cometidos por el sujeto y calculando la probabilidad condicionada, $P(x_{ij} \mid x_{.j})$, que es la probabilidad de que el sujeto cometa exactamente x_{ij} errores en el texto i, condicionada a que su número total de errores ha sido $x_{.j}$, se puede probar que esta distribución no depende del parámetro θ. Este importante resultado es lo que llevó a Rasch a considerar que, si el objetivo era la medición del parámetro θ_j a partir del número de errores observados (x_{ij}, x_{kj}), lo podía realizar utilizando como estimador el número total de errores $x_{.j}$, siendo este un estimador suficiente para el parámetro θ (Rasch, 1960; Santisteban, 1990, p.246).

En la literatura se pueden encontrar múltiples referencias en las que de una forma u otra se pone de manifiesto la gran importancia de la aportación de Rasch a la medición en Psicología. Benjamín D. Wright, en la obra editada por Embretson y Hershberger en 1999, dice que Rasch (1960, 1961, 1963, 1968, 1969, 1977) muestra que la formulación usando una familia compuesta de Poisson (Bernouilli o binomial), no solo es suficiente para la construcción de mediciones *estables*, sino que el uso de las propiedades multiplicativas de las leyes de Poisson es la única solución matemática para tener funciones medibles en teoría de tests. Andrich (1995, 1996) confirma que en el modelo de Rasch la *separabilidad* requiere la distribución de Poisson para estimar medidas a partir de observaciones discretas y Bookstein (1996), por su parte, también muestra la necesidad del uso de la distribución compuesta de Poisson, en cuanto a que se satisfacen los requisitos de *divisibilidad,* definidos por Levy (1937) y por Kolmogorov (1950).

En resumen, la propuesta de Rasch implica no sólo que se pueden obtener estimadores suficientes de los parámetros, con medidas directas, y haciendo uso de las propiedades de las distribuciones y los procesos de Poisson compuestos, sino también que las mediciones que se hacen cumplen los requisitos teóricamente exigibles, como son la *estabilidad*, *separabilidad*, *independencia muestral*, así como el requisito de *aditividad conjunta* que, en teoría de la medición, es un requisito fundamental. También en el modelo TRI conocido como modelo de Rasch se mantienen todas estas propiedades.

En los modelos de Rasch, el requisito de la aditividad conjunta se construye aplicando la probabilidad inversa a los datos empíricos y contrastando la bondad del ajuste de esos datos a esa medida conjunta (Fischer, 1968). Al respecto, Brogden (1977) dice que el modelo de Rasch es un caso especial de *medición aditiva conjunta* y que un ajuste del modelo de Rasch implica que se satisface el axioma de cancelación.

En los años ochenta del pasado siglo aparece en la literatura un nuevo concepto, acuñado por Roskam (1983), que es el de objetividad específica. En el modelo de Rasch también se cumple el principio de *objetividad específica*, o equivalentemente el de *independencia muestral*, que requiere que los parámetros de los sujetos se estimen independientemente de los de los parámetros de los ítems y viceversa.

La medición de los sujetos es, en un sentido estocástico, independiente del conjunto específico de ítems seleccionados del universo de contenidos que definen el dominio específico de comportamiento. Una consecuencia de esto es que un modelo de medición con objetividad específica es el único instrumento que tiene sentido para la definición y la medición de constructos teóricos del comportamiento que se supone que subyace en el dominio. En palabras de Roskam y Jansen (1984), la objetividad específica, y por ende el modelo de Rasch, es una condición metodológica básica para la definición y medición de constructos teóricos como, por ejemplo: aptitud verbal, control interno, maquiavelismo. Es de especial importancia para el estudio de las actitudes y de los rasgos de personalidad, donde puede ser muy útil aislando y midiendo tales constructos.

3. La validez de los tests psicométricos

La validez está considerada como el requisito previo y el aspecto fundamental para el uso y aplicación de cualquier instrumento de medida psicométrico y aunque parece obvio que cualquier juicio sobre validez implica y depende del modelo de medida adoptado, curiosamente la teoría de la validez ha seguido, durante mucho tiempo, una evolución total o parcialmente disociada de los desarrollos de la teoría de la medición. Afortunadamente, en los últimos años se observa un movimiento constante y creciente de convergencia, tanto desde la teoría de la medida como desde la validez, que conduce necesariamente a una teoría unificada que permitirá dar respuesta a muchos de los problemas que plantea una adecuada evaluación psicológica.

La concepción actual de la validez va más allá de una mera reformulación y supone un importante cambio filosófico, desde una orientación puramente pragmática y empirista, que caracterizó a la Psicometría en la primera mitad del siglo XX, a un nuevo y creciente interés en la teoría psicológica (Angoff, 1988). Este cambio conceptual ha seguido dos trayectorias en paralelo (Messick, 1989): (a) Desde el establecimiento de numerosos criterios específicos de validez, pasando por unos pocos tipos de validez, hasta una concepción de validez unitaria y (b) Desde la predicción hasta la explicación como aspecto sustantivo de la validez, de forma que la utilidad, la relevancia, o la importancia de la predicción no se puede actualmente evaluar en ausencia de una sólida interpretación de las puntuaciones que se obtienen al aplicar los tests.

Entre los primeros acercamientos empíricos al problema de la validez, podemos citar el procedimiento de evaluación de los ítems utilizado por Binet y Simon (1908). Estos autores, procuraban elegir los mejores ítems utilizando el criterio de diferenciación por edad, criterio que se sustenta en la observación procedente de la Psicología evolutiva de que el niño va desarrollando sus aptitudes intelectivas a medida que crece.

Geisinger (1992) atribuye a Clark Hull la aportación de que la correlación entre pruebas es un procedimiento adecuado para conocer si un test es o no valido. En 1928 Hull se limitó a sugerir, sin llegar a explicitar, la que finalmente se convertiría en la definición de coeficiente de validez ρ_{XY}, entendido como la correlación, en valor absoluto, entre un predictor X y un determinado criterio Y (Gulliksen, 1950a).

Con las primeras definiciones de validez, pronto surgirá la problemática de la atenuación de la medida. Las correlaciones entre variables no son todo lo altas que cabría desear, lo que podría explicarse porque los errores de medida están afectando o atenuando la autentica correlación entre los rasgos (ver Santisteban, 1990 pags. 154-160; Santisteban y Alvarado, 2001 pags.53-55). Para dar respuesta a esta problemática Cureton (1950) incluirá la atenuación en sus definiciones de la validez, entendiendo la validez como la correlación entre las puntuaciones observadas en el test y las puntuaciones verdaderas del criterio. Además, distinguirá entre la validez del test y su poder predictivo, definiendo este último como la correlación entre puntuaciones observadas en el test y las puntuaciones observadas en el criterio; y distinguirá ambas de lo que denominó relevancia o correlación entre las puntuaciones verdaderas, tanto del predictor como del criterio. Sin embargo, las propuestas de Cureton (1950) no serán asumidas por la comunidad científica, aceptándose en su lugar la definición de coeficiente de validez y las consideraciones realizadas por Gulliksen (1950a). Para Gulliksen no se puede hablar de una única validez del test, existiendo tantas como coeficientes de validez. Por ejemplo, el *ACE Psychological Examination* tiene una validez para predecir el nivel en Inglés y otra para predecir el nivel en Latín. La validez también cambia de un colegio a otro y de una ocasión a otra. En otras palabras, la validez no puede ser considerada como una característica fija o unitaria de un determinado test. Habrá tantos coeficientes de validez como usos se hagan de él y, por lo tanto, sólo es correcto hablar de la validez de un test cuando hace referencia a la predicción de un criterio específico.

El pensamiento mayoritario acerca de la validez en la primera mitad del siglo XX queda perfectamente ejemplificado en Anastasi (1950, p.67): "*un test sólo puede ser objetivamente validado como medida de un determinado criterio... Afirmar que con un test se mide algo distinto a su criterio es pura especulación.*"

En una de las pocas disensiones de la concepción mayoritariamente aceptada de validez, Rulon (1946) defiende un concepto alternativo que se aplicaría a los tests de rendimiento académico. Argumenta que esos tests no requieren de un criterio externo para ser validados (constituyen su propio criterio) puesto que son válidos desde una perspectiva racional: en los tests de rendimiento académico la exigencia de mostrar su validez se logra mediante la revisión del test por expertos en la materia y la verificación de que su contenido representa una muestra satisfactoria de su dominio.

Considérese que esa época estaba todavía inmersa en la tradición conductista en Psicología e imbuida del marco filosófico del positivismo lógico, en la que aspectos como *significado* o *categoría de respuesta,* eran considerados inaceptables por su *subjetividad* para construir una *verdadera* ciencia de la conducta. El pensamiento dominante acerca de los conceptos en Psicología se resumía en que estos eran: (a) *constructos* explicativos ficticios e inventados o construidos a conveniencia por los científicos de la conducta (b) *inobservables*, aunque *inferidos de la respuesta*, es decir, objetivamente inferidos (no interpretados) a partir de registros de conducta, esto es, a partir de las respuestas sin que medie interpretación por parte del investigador.

En este clima filosófico tiende a prevalecer la visión de que la validez, o la validación de las puntuaciones del test, consiste en su capacidad para predecir el *criterio* observable de conducta. Por lo tanto, todo test debería validarse mediante procedimientos que permitieran mostrar sus relaciones con algún *resultado* medido externamente. En toda validación se exige proporcionar ejemplos en los que se muestre la relación entre las puntuaciones del test con la realización o desempeño en programas prácticos de entrenamiento, tales como un grupo con alguna patología diagnosticada por el test, en comparación con un grupo control. De este modo, podríamos evaluar si un test tiene *validez predictiva* respecto a un criterio de conducta específico (generalmente serán múltiples las conductas estudiadas y por lo tanto obtendremos múltiples coeficientes de validez). En este mismo espíritu positivista se acuñó el término *validez concurrente*, como aquella en la que se toman las medidas del test predictor y del test criterio en el mismo momento, y que serviría para justificar la sustitución de unos instrumentos por otros.

A finales de la década de los 40 y comienzos de los años 50 del siglo XX, con los avances en metodología estadística y el comienzo de la aplicación de técnicas de análisis de los ítems respecto a la puntuación total del test o al criterio de medida externo, surgen distintos términos para la validez, entre los que se vislumbra ya la idea que dará lugar a la validez de constructo:

- Guilford (1946) introduce el concepto de validez factorial, entendida como la que proporcionan las comunalidades o pesos factoriales más significativos. Este tipo de validez nos proporciona información acerca de qué es lo que realmente mide el test y la respuesta se podrá dar en términos de factores y pesos. Como veremos más adelante, Guilford estaría anticipando una de las técnicas más importantes para la validación de constructo.
- Cronbach (1949) distingue entre la validez lógica y la empírica, discutiendo en su último capítulo los problemas de la validez de contenido en los tests educativos. La validez lógica se basaría en analizar qué mide el test, mientras que la empírica se centraría en el estudio de la correlación entre las puntuaciones del test y otra medida, generalmente un criterio, que se conoce o supone que cubre algún aspecto de interés.
- Gulliksen (1950b) introduce el concepto de *validez intrínseca* que hace referencia a la existencia de factores explicativos, que se descubren al examinar los datos de las distintas medidas y estudiar los patrones de covarianza. Obsérvense las similitudes entre el concepto de validez intrínseca y el de validez factorial.
- Anastasi (1954) divide la validez en *aparente, de contenido, factorial y empírica*. Aunque la validez aparente no será considerada como un verdadero tipo de validez sino, simplemente, como una mera estrategia del psicómetra.

Con la intención de introducir orden en el estado caótico en los procedimientos de construcción de los tests, en 1954 la *American Psychological Association* (APA) en colaboración con la *Educational Research Association* (AERA) y el *National Council on Measurement in Education* (NCME) publicaron las recomendaciones técnicas para tests psicológicos y otras técnicas diagnósticas. En la preparación de estas *primeras normas* intervinieron de forma activa Cronbach y Meehl como jefe y miembro del comité responsable, respectivamente, apostando por un nuevo tipo de validez, la validez de constructo, como quedó reflejado en las recomendaciones publicadas por la APA (1954) y en un artículo posterior (Cronbach y Meehl, 1955), a pesar de que todavía el pensamiento de la época estaba dominado por el positivismo lógico.

La validación de constructo fue concebida para situaciones excepcionales como, por ejemplo, cuando se necesita interpretar las puntuaciones del test como medida de un atributo o cualidad que no es posible definir de otro modo, o no existen criterios válidos, o no hay universos de contenido que se puedan aceptar como adecuados para definir la cualidad a medir. De hecho, Cronbach y Meehl (1955) afirman que la validación de constructo se introduce para aquellas investigaciones específicas, que requieran del desarrollo de tests, en las que las perspectivas convencionales de validación no sean apropiadas.

En las recomendaciones técnicas (APA, 1954) se proponen cuatro tipos de validez: la concurrente, la predictiva, la de contenido y la de constructo. El peso fundamental lo seguirá teniendo la *validez empírica* (concurrente y predictiva), defendiéndose que el test deberá validarse en función del uso que se le pretenda dar. No obstante, en este documento hay un conjunto de cambios respecto al pensamiento precedente que merece la pena señalar:

a) Aparece como categoría nueva la *validez de constructo*, que ya había sido anticipada en otras denominaciones como la validez lógica, la validez factorial o la validez intrínseca.

b) La validación pasa, de ser una responsabilidad propia del editor, a ser compartida por los expertos que utilicen el test.

c) Se especifica la exigencia de indicar la utilidad del test, de manera que pueda ser validado según la estrategia más adecuada.

La novedad más importante es la inclusión de la validez de constructo. Un constructo sería admisible siempre que se enmarcara en una red nomológica, y ello exige establecer leyes estadísticas y/o determinísticas que liguen unas propiedades observables con otras, los constructos teóricos con los observables, y los constructos teóricos entre sí. En esta definición inicial se aprecia cómo la racionalidad y la metodología se desprenden de la filosofía positivista de la ciencia dominante en su tiempo. La fundamentación epistemológica del positivismo descansaba en la suposición de que la teoría, y las observaciones que permiten poner a prueba la teoría, eran independientes, y que la verdad de las observaciones no era problemática. Desde esta perspectiva se defendía que se debería de especificar completamente el constructo, su red nomológica, y que la tarea de validar el constructo consistía *simplemente* en verificar las relaciones hipotetizadas.

Las revisiones a las recomendaciones técnicas que se hacen en los años 1966 y 1974 pasaron a denominarse Normas (S*tandards*) para tests educativos y psicológicos (APA, AERA y NCME, 1966, 1974). La validez concurrente y predictiva se integrarán en la validez relativa a un criterio, dando lugar a los tres tipos de validez clásicos: contenido, criterio y constructo. Sin embargo, con el tiempo, la validez de constructo se va a ir imponiendo hasta convertirse en el aspecto que aglutina a las distintas fuentes de validación.

Desde que la validez de constructo se describió por primera vez, el propio Cronbach, así como otros autores (e.g. Campbell, 1960) matizarán que era poco realista tratar de especificar la red nomológica completa en situaciones reales o aplicadas. Esta limitación, aunque de naturaleza pragmática, también refleja un cambio filosófico respecto al anterior marco positivista aplicado a las ciencias sociales. Cronbach (1975) reconocerá que los planteamientos que inicialmente defendió no son factibles debido a que el desarrollo en las ciencias sociales no ha sido el esperado, además se observa que las generalizaciones sobre las que se construyen las redes nomológicas decaen con el paso del tiempo, y las interacciones de mayor orden difícilmente se sostienen. Más recientemente, Cronbach (1989) dirá que era pretencioso emplear un lenguaje positivista en una ciencia inmadura como la Psicología y que es mucho más razonable plantear programas de validación más limitados que contemplen la relación entre unos pocos constructos, para posteriormente ir ampliando estas relaciones en la medida en que se vayan cumpliendo los objetivos propuestos.

Embretson (1983) publica un trabajo de enorme valor, que todavía hoy no ha sido totalmente entendido ni asimilado por parte de la comunidad científica (ver Schwager, 1991). En este artículo se explicitan las dos dimensiones que es necesario abordar en todo intento de validar el constructo que se pretende medir, al utilizar un test determinado: la representación del constructo y los aspectos nomotéticos.

Embretson (1983) afirma que si analizamos las primeras formulaciones de la validez de constructo y las comparamos con las producidas diez años antes, observamos que se ha producido un cambio paradigmático en la teoría psicológica con la irrupción de la Psicología cognitiva: el objetivo de la Psicología teórica ha cambiado desde la pretensión de explicar las relaciones antecedentes-consecuentes (estructuralismo), a intentar explicar el funcionamiento de los procesos que conforman el sistema y de los subsistemas subyacentes (funcionalismo). Como todo cambio paradigmático, la perspectiva de la Psicología cognitiva y específicamente la centrada en el análisis del procesamiento de la información supone cambios, no solo en las cuestiones planteadas, sino también sobre los datos que se consideran relevantes. Puesto que la validación del constructo fue concebida por Cronbach de modo similar al de la construcción de las teorías científicas, el cambio de paradigma al que estamos asistiendo debería de tener una notable influencia, no sólo en las investigaciones sobre validación de constructo, sino en toda investigación psicológica. Es decir, las cuestiones básicas y los métodos apropiados para determinar los constructos que permitirían dar cuenta de la varianza de las puntuaciones obtenidas al aplicar los tests, deben de sufrir un cambio radical para adaptarse al nuevo paradigma del procesamiento de la información. Ahora, ya no bastará con establecer las redes de relaciones entre conceptos, sino que será necesario como requisito previo que el constructo esté adecuadamente representado.

La representación del constructo tiene que ver con la identificación de los mecanismos teóricos que subyacen a la realización de la tarea. El objetivo de las investigaciones de representación del constructo es la descomposición de las tareas en procesos subyacentes. En el nuevo paradigma del procesamiento de la información, la representación del constructo se refiere a la dependencia de las respuestas a una tarea respecto a los procesos, estrategias y conocimientos que están involucrados en su realización. Es decir, se trata de modelar lo que subyace a la realización de la tarea que se le pide al sujeto (estas tareas pueden ser bien variadas), en términos de los procesos, estrategias y conocimientos involucrados en su realización. Téngase en cuenta que estas variables subyacentes pueden proceder no solo del ámbito cognitivo, sino también del afectivo, e incluso del social.

Embretson (1983) observa que los constructos que son identificados en la descomposición de tareas, no tienen necesariamente implicaciones para las diferencias individuales. Es decir, un proceso o una estrategia puede ser un aspecto esencial de la realización de la tarea, pero la población estudiada puede no variar sistemáticamente en su habilidad para realizarla. La investigación de la representación de constructo tiene que ver con la variabilidad de la tarea más que con la variabilidad de los sujetos.

La perspectiva nomotética o clásica, se refiere a la red de relaciones que se establece entre un test y otras medidas del constructo. El enfoque nomotético se apoya en la fuerza, la frecuencia, y los patrones de relaciones significativas con otras medidas, como son otros rasgos, medidas de criterio, etc. El enfoque nomotético, a diferencia de la representación del constructo, indica la importancia del test como una medida de las diferencias individuales.

Cuando Cronbach y Meehl (1955) enuncian la validez de constructo, el énfasis principal en Psicología Experimental era el de las relaciones antecedentes- consecuentes, con poco interés en los mecanismos intervinientes que eran relevantes para la representación del constructo. Aunque los investigadores *correlacionales*, que utilizaban prioritariamente las técnicas de análisis factorial, estaban interesados en la descomposición de tareas, la representación del constructo se confundía completamente con el enfoque nomotético. Es decir, los *componentes* que surgen del análisis factorial se basan en las correlaciones entre tareas, pero como las correlaciones entre tareas reflejan múltiples influencias, incluyendo prerrequisitos de conocimiento, comunalidades educativas, y comunalidades genéticas, así como componentes teóricos subyacentes comunes, el análisis factorial era incapaz de separar y discernir cuáles eran los componentes o aspectos subyacentes a las tareas.

En el enfoque propuesto por Embretson las especificaciones del test se construyen desde la teoría, y posteriormente, las implicaciones de estas especificaciones se estudian empíricamente. De este modo, la representación del constructo es una fase de investigación del proceso de validación del constructo. Las especificaciones realizadas sobre cada ítem son consideradas como teoría de una tarea, y la teoría se pone a prueba mediante modelos matemáticos que intentan explicar los constructos. Posteriormente, mediante la estimación de las propiedades cuantitativas de los ítems con respecto a los constructos teóricos, durante el desarrollo del test, pueden descartarse los ítems que no reflejen las propiedades teóricas deseadas en el test que se va a desarrollar.

La fase nomotética indica aspectos que sólo se pueden estudiar cuando se procede a partir de constructos teóricos definidos y representados en la anterior fase. La investigación nomotética debe de considerarse como una operación convergente para evaluar la teoría que está representada en el constructo.

En la formulación de Embretson, la representación y el enfoque nomotético son fases interactivas del proceso de validación del constructo, que encontraría su ámbito idóneo de aplicación en los modelos componenciales de la TRI como, por ejemplo, el modelo logístico lineal de Fischer (1973) o el modelo multicomponente de rasgo latente de Withely (1980).

La edición de las *Normas* que durante más tiempo ha estado vigente, hasta el momento, ha sido la de 1985 (APA, AERA y NCME, 1985). En estas *Normas* se produce un importante cambio al intentar dar un tratamiento unificado al proceso de validación. En esta edición se establece que la validez es un concepto unitario que se refiere a la adecuación, significado y utilidad de las inferencias específicas realizadas a partir de las puntuaciones obtenidas en los tests. Los tres tipos clásicos de validez (contenido, criterio y constructo) pasan a ser solamente distintas fuentes o evidencias de validez. Desde que Loevinger (1957) afirmara que toda la validación lo es del constructo, esta idea ha ido afianzándose progresivamente entre los teóricos de la validez, siendo bastante generalizada a principios de los 80. Sin embargo, debemos de tener en cuenta, que el consenso acerca de este concepto unificado de validez, es mucho más amplio entre psicólogos que entre educadores y, puesto que las normas las hacen conjuntamente entre ambos, es razonable que al elaborar las normas exista en este punto cierta ambigüedad.

En las normas de 1985, tal como apunta Anastasi (1988), la validez relativa a un constructo se derivará del establecimiento de interrelaciones entre distintas medidas de conducta y las puntuaciones del test que se pretende validar. La validación es un proceso que se realiza mediante la acumulación gradual de información procedente de distintas fuentes. Se deberá de tener en cuenta en el proceso todo dato que permita esclarecer la naturaleza del constructo bajo estudio y de las condiciones que afectan a su desarrollo y manifestaciones. La validación de constructo rara vez podrá consistir en una simple respuesta, afirmativa o negativa, de si una determinada interpretación está justificada. En general, los resultados de un determinado estudio o línea de investigación darán como resultado la modificación del test, del constructo, del marco conceptual en que se enmarca el constructo, o de todo ello. Así, la validación de constructo se convierte en un aspecto fundamental en la construcción del test. Y, como dice Anastasi (1986), casi cualquier información recogida en el proceso de desarrollo o uso del test es relevante en el estudio de su validez.

Messick (1989; 1995) considera que la definición de la validez que se desprende de las *Normas* de 1985 es fragmentaría e incompleta, defendiendo un nuevo concepto unificado de validez que incluya toda posible fuente de evidencia que sirva al propósito de validación, incluyendo los valores y las consecuencias sociales del uso de los tests. Se concibe la validez de constructo como un concepto unificado que integra las consideraciones acerca del contenido, del criterio y las consecuencias de la aplicación del instrumento. Es decir, en un marco de constructo se realizará la evaluación empírica de las hipótesis acerca del significado de las puntuaciones del test y de las relaciones teóricamente relevantes que incluirán las de naturaleza científica y aplicada.

Para Messick (1989) la validez es un juicio evaluativo global del grado en el que la evidencia empírica y los razonamientos teóricos apoyan lo adecuado y lo apropiado de las interpretaciones y acciones basadas en las puntuaciones de los tests o en otros modos de evaluación. La validez no sería una propiedad del test, sino que dado que el proceso de validación se puede concebir como una investigación acerca del significado de las puntuaciones del test, lo que se valida en realidad es el uso e interpretación que se realiza de las puntuaciones para una aplicación concreta. Pero las puntuaciones no son únicamente función de los ítems o de las condiciones estimulares, sino también de las personas que responden y del contexto de evaluación. En particular, lo que se examina en la validación es el significado y la interpretación de las puntuaciones, así como las acciones que pudieran afectar al resultado (Cronbach, 1971). Al extender el significado de las puntuaciones a las acciones, tanto para los sujetos y grupos, como a las situaciones y contextos, la validación se entiende como un proceso continuo y dinámico que no tiene fin.

Los principios de la validez se aplican, no sólo a las inferencias derivadas de las puntuaciones del test, tal como han sido ordinariamente concebidas, sino también a las inferencias basadas en cualquier medio de observación o documentación consistente con las conductas o los atributos estudiados. El término *puntuación* toma un sentido genérico que significa cualquier codificación de observaciones consistentes o regularidades en la realización del test, cuestionario, procedimiento de observación u otras formas de evaluación.

Seis serian los aspectos de la validez de constructo que habría que poner de relieve para situar el concepto central de validez, los cuales funcionan como un criterio general de validez para la medición en Educación y en Psicología (Messick, 1995). Esos aspectos son:

(1) El de contenido, que incluye la relevancia del contenido, la representatividad del dominio, y la calidad técnica del instrumento.

(2) El sustantivo, que se refiere a las justificaciones teóricas de las respuestas que de forma consistente se observan al aplicar el test, incluyendo el modelado de procesos de la tarea realizada (Embretson, 1983), junto con la evidencia empírica de que los procesos teóricos están realmente engarzados con las tareas que realizan los sujetos.

(3) El estructural, que evalúa la fidelidad de la estructura de puntuaciones respecto a la estructura del dominio del constructo (Loevinger, 1957; Messick, 1989).

(4) La generalización o "generalizability" que consiste en examinar la extensión para la cual las propiedades e interpretaciones se generalizan a otros grupos, situaciones y tareas (Cook y Campbell, 1979; Cronbach, et. al. 1972), incluyendo la generalización de la validez para los tests relacionados con el criterio (Hunter, Schmidt, y Jackson, 1982).

(5) El externo, que incluye el establecimiento de relaciones convergentes y discriminantes mediante las comparaciones multimétodo-multirrasgo (Campbell y Fiske, 1959), así como los aspectos de relevancia de criterio y de utilidad práctica (Cronbach y Gleser, 1965).

(6) Las consecuencias, aspecto que evalúa el valor de las implicaciones de las interpretaciones de las puntuaciones como una base para la acción, tanto como las consecuencias reales y potenciales del uso de los tests, especialmente respecto a las fuentes de no-validez relacionadas con los temas de sesgo, equidad, y justicia distributiva (Messick, 1980; 1989).

Se han de evaluar las consecuencias de lo pretendido, y de lo no pretendido, con la interpretación de las puntuaciones y el uso, tanto a corto como a largo plazo, del test. Por ejemplo, en el ámbito de lo educativo, puesto que la evaluación del rendimiento promete un beneficio potencial para el aprendizaje y la enseñanza, es importante acumular evidencia de estas consecuencias positivas, tanto como de que las negativas no son relevantes.

En cuanto a las consecuencias adversas de la aplicación del test, es fundamental asegurarse que ningún impacto negativo en los individuos o grupos se derive de alguna fuente de no-validez del test como, por ejemplo, la utilización de constructos inadecuadamente representados en los ítems del test, o que una parte de la varianza empírica que se obtiene al aplicar el test sea irrelevante para la medida del constructo (Messick, 1989). Es decir, nunca una puntuación baja en el test debería de darse como consecuencia de que el test no contemple aspectos relevantes del constructo, que de estar presentes, permitirían a las personas afectadas mostrar su verdadero nivel de competencia. Por otra parte, las puntuaciones bajas no deberían de explicarse por una contaminación en la medida de aspectos irrelevantes que interfieran en la demostración de la competencia real de algunas personas.

El análisis de estos seis puntos permitirá formular un juicio global de validez. La esencia del concepto unificado de validez es que lo apropiado, lo significativo, y lo útil de las inferencias que se basan en las puntuaciones del test, son aspectos inseparables y que el poder de integración se deriva de la fundamentación empírica de su interpretación. Tanto el significado como los valores se han de integrar en este concepto de validez.

Las *Normas* de 1999 (APA, AERA y NCME, 1999) recogen fundamentalmente las ideas vertidas por Messick en la última década. En ellas se define la validez como una evaluación global del grado en el que la teoría y la evidencia apoyan las interpretaciones de las puntuaciones que se hacen en cada uso específico de los tests. Por lo tanto, la validez es la consideración más importante en el desarrollo y evaluación del test. El proceso de evaluación consiste en obtener una justificación suficiente que apoye las interpretaciones que se hacen de las puntuaciones del test. Para ello, debe de prestarse especial atención a los aspectos que se mencionan en los siguientes apartados:

a) La descripción de la interpretación que se proponga, proporcionará el marco conceptual en el que especificará el significado del constructo, aclarando los propósitos de la evaluación, distinguiendo el constructo de otros, delimitando los conocimientos y procesos implicados en el constructo, especificando los contextos en los que se debe de utilizar el test, e indicando cómo las distintas medidas del constructo se relacionan con otras variables.

b) Las decisiones sobre qué justificaciones son pertinentes en el proceso de validación, deben de clarificarse mediante un conjunto de proposiciones, que apoyen la interpretación propuesta para el propósito particular de la evaluación.

c) La validación es un proceso que exige la comprobación de distintas proposiciones, por lo que el encontrar apoyo en unas, no elimina la necesidad de poner a prueba las demás. Por ejemplo, en selección de personal, encontrar una fuerte relación entre el predictor y el criterio no será suficiente para justificar el uso del test. Además, será necesario mostrar la importancia y adecuación del criterio que se mide.

d) La validación es una responsabilidad compartida entre el que desarrolla el test y el que lo utiliza. El que lo desarrolla debe de racionalizar y proporcionar justificaciones relevantes en apoyo del uso que se propone para el test, pero el que lo utiliza es responsable en último término de decidir si en esa situación particular está o no justificado su uso.

En las *Norma*s de 1999 se sostiene que la validez es un concepto unitario que supone la evaluación integral del grado en que la acumulación de distintos hallazgos permiten apoyar la interpretación propuesta para el test, definiéndose las siguientes fuentes de evidencia que se basan en:

a) *El contenido del test.* Incluye análisis lógicos y empíricos respecto a la relevancia y representatividad del contenido, con respecto al dominio en el que el test está definido y a la interpretación que se propone para las puntuaciones del test.

b) *Los procesos de respuesta.* Análisis teóricos y empíricos de los procesos de respuesta en los examinados que justifiquen la relación entre el constructo y las respuestas.

c) *La estructura interna.* Los análisis de la estructura interna del test pueden indicar el grado de relación entre los ítems del test y los componentes que conforman el constructo sobre el cual la interpretación del test se fundamenta.

d) *Las relaciones con otras variables.* Los análisis de las relaciones de las puntuaciones del test con variables externas puede suponer una fuente de justificación muy importante para la validez. Las variables externas pueden incluir medidas de algún criterio que se espera que prediga el test, así como las relaciones con otros tests que se hipotetiza miden el mismo constructo, y test que miden constructos relacionados o diferentes. Se trata de averiguar si las relaciones son consistentes con el constructo subyacente propuesto. En esta fuente de información se incluyen:

- *Evidencia de convergencia o de discriminación*: Las relaciones entre medidas de constructos similares, proporcionan información convergente, mientras que las relaciones con medidas de constructos diferentes, proporcionan información sobre la discriminación entre ellos. Las relaciones entre diferentes métodos de medida puede ayudar a afinar y elaborar el significado e interpretación de la puntuación. En la literatura se proponen métodos correlacionales que ponen en relación distintos rasgos y medidos por distintos métodos, aportando evidencia de validez convergente, altas correlaciones entre el mismo rasgo, aún estando medido por diferentes métodos y discriminante, bajas correlaciones entre rasgos, aún medidos por el mismo método.

- *Relaciones entre el test y un criterio*. Para una determinada situación de evaluación se determina en un estudio test-criterio el valor del test para la predicción del criterio analizando su relevancia, fiabilidad y validez predictiva.

- *La Generalización de la validez*. Un aspecto importante en el ámbito educativo y laboral es el grado en que la validez, en relación con un criterio, se puede generalizar a nuevas situaciones sin la necesidad de realizar estudios adicionales para estas nuevas situaciones. Hay que ser extremadamente exigentes y rigurosos en cuanto al cumplimiento de todos los requisitos metodológicos que conllevan los procesos de inferencia. Recientemente, los procedimientos meta-analíticos han mostrado que, en algunos dominios, gran parte de la variabilidad encontrada en los estudios iniciales de generalización se debía a artefactos estadísticos tales como fluctuaciones de las muestras, restricción del rango o escasa fiabilidad en los criterios.

e) *Evidencia basada en las consecuencias de la evaluación*. En los últimos años se han incluido las consecuencias (tanto intencionadas como no intencionadas) de la evaluación como aspectos que afectan a la validez del instrumento.

La validez debe de estar presente desde el principio del proceso de construcción del test, en vez de estar limitada a las últimas etapas de su desarrollo como ocurría en la visión tradicional de la validez.

El proceso de validación comienza con una definición clara y precisa del constructo a medir, derivada de una determinada teoría psicológica, o del consenso entre las existentes. Se deben de establecer las relaciones, tanto convergentes como discriminantes, con respecto a otros constructos relevantes en esa teoría. Posteriormente, se comienza la elaboración de los ítems de modo que sus enunciados se ajusten a la definición del constructo, para luego realizar análisis empíricos que permitan seleccionar los más adecuados a partir del conjunto inicial de ítems (Santisteban, 1990). Estos análisis posteriormente deben de completarse con otros que permitan conocer mejor la estructura interna del test (análisis factoriales exploratorios y confirmatorios, análisis discriminantes, análisis de cluster, etc.). Finalmente, tal y como propone Anastasi (1986) se deben de realizar otros estudios adicionales en los que se pongan a prueba las interpretaciones de las puntuaciones del test.

Este "nuevo" acercamiento que ya fue defendido por Jackson (1970), contrasta fuertemente con la vieja idea de primero construir un instrumento procurando que sea muy fiable, y luego, en la etapa final, someterlo a estudios de validación. Ahora, la validación ya no es la etapa final de un proceso de construcción y en la que apenas interesa aquello que no esté relacionado con la predicción del criterio de interés, sino que es el aspecto fundamental y, por lo tanto, guía la construcción del test desde principio a fin.

Observamos que aunque continúa vigente la perspectiva empirista de la validación del constructo, ya no es un empirismo ciego, puesto que se exige una teoría o marco teórico de referencia. Se debe de tener un especial cuidado en el desarrollo de las definiciones del constructo que se pretende medir. Este proceso debe de ser revisado por expertos que juzguen si, efectivamente, los ítems que se han generado son consistentes con las especificaciones derivadas de la teoría, que es la que debe de guiar la construcción del test. Todos estos procedimientos de verificación preceden y deben de integrarse con el resto de los procedimientos formales y empíricos, de modo que el proceso de construcción del test suponga la puesta a prueba de hipótesis, que han de ser contrastadas durante el proceso de validación del constructo.

La Psicología cognitiva, como marco de referencia para la validación de los constructos, ha ido ganando con el tiempo mayor importancia, siendo hoy día la que fundamentalmente nutre del sustrato teórico en la explicación de cómo los individuos aprenden y procesan la información. Este progreso se produce sobre todo desde que Embretson (1983) la consideró y la propuso como la fuente de conocimiento que permitiría representar las distintas tareas que llevan a una realización particular. En estos últimos años, la Psicología cognitiva en general, y los modelos de procesamiento de la información en particular, han avanzado de forma importante, produciendo poderosas técnicas experimentales y cuantitativas de descomposición de tareas, con lo que la propuesta de Embretson se ha revelado acertada, incorporándose a las Normas, bajo la denominación de la *representación del constructo*. No obstante, queda por hacer lo más difícil. La propuesta de Embretson no se quedaba en utilizar la descomposición de tareas para hacer una mejor descripción del constructo, sino que iba más lejos, proponiendo que la construcción de los tests se realice mediante el uso de los modelos componenciales de la teoría de la respuesta al ítem, para así lograrse una adecuada modelización de los constructos subyacentes, con la consiguiente ganancia en validez.

NOCIONES DE TEORÍA DE LA MEDIDA

1. Métrica y Espacios Métricos

1.1. Introducción

En esta primera parte se trata de introducir las nociones más básicas de la Teoría de la Medida, pues su conocimiento es necesario para toda aproximación a la investigación de cualquier fenómeno, desde una aproximación con enfoque probabilista y estadístico, así como al de la validez de los resultados.

La Teoría de la Medida abarca muchos campos, desde el filosófico, hasta operaciones matemáticas de naturaleza compleja, como pueden ser las integrales iteradas y los espacios producto infinitos.

Al estar esta obra dirigida fundamentalmente a quienes realizan estudios de investigación en Psicología, en Ciencias Sociales y en Ciencias de la Salud, lo que se pretende es introducir las nociones y los conceptos básicos, tratándolos de forma general, sin utilizar instrumentos matemáticos avanzados. El objetivo es que el investigador adquiera las nociones y el lenguaje propios de esta materia, para buscarle el significado en su campo, sin entrar en demostraciones tediosas, incluso tan complicadas o abstractas que, en lugar de llevarle a un mayor conocimiento, lo pueden llevar a confusión. Puesto que muchas veces esto no es posible en un tema tan complejo y el propósito primordial de esta obra es didáctico y esta dirigido a los que inician su andadura en la investigación, por una parte, tratamos de no repetir la estructura y algunos de los contenidos que están bien tratados en los manuales, por otra, antes de trivializar o de perder rigurosidad, remitimos al lector a los textos y/u otras fuentes donde se trate el tema o problema concreto con la debida extensión, rigurosidad y profundidad.

En la Teoría de la Medida y, por ende, en la medición psicológica de los atributos, todos los conceptos que se utilizan en el quehacer científico se pueden incluir en dos clases muy amplias, los que se han dado en llamar *cualitativos* y los *cuantitativos*. Los conceptos cualitativos son aquellos que son comparativos, o simplemente clasificatorios. Los conceptos cuantitativos son los *mensurales* y/o *métricos*. Es decir, que al menos sean susceptibles de medida, siendo métricos cuando entre esas medidas se puedan establecer un cierto tipo de red de relaciones.

La clasificación de los conceptos científicos en una u otra clase, es casi siempre un problema epistemológico, si bien, no podemos negar que en la naturaleza o en la realidad que estemos investigando, existen propiedades que hacen que el fenómeno sujeto de estudio, haya que considerarlo esencialmente cualitativo, o bien, esencialmente cuantitativo. Hay que tener en cuenta, que hacer esa clasificación depende tanto de la naturaleza del fenómeno, esto es, del fenómeno del mundo real que se está investigando, como de la atribución de cualitativo o cuantitativo que le hace el investigador, basándose en la estructura conceptual desde la que aborda el problema. Es decir, consideramos que al no ser el mundo real, ni cualquier fenómeno en particular, de naturaleza puramente cualitativa o cuantitativa; al no ser ningún fenómeno cualitativo o cuantitativo *per se*, es el investigador quien tiene una gran responsabilidad de incluirlo en una u otra clase, al hacer su aproximación al estudio de ese fenómeno (Santisteban, 2003). Obviamente, no estamos aquí considerando las opiniones, ni de aquellos que consideran que los atributos no físicos, y menos aún las variables psicológicas, sean susceptibles de medida y, por ende, adecuadas para hacer análisis cuantitativos de los fenómenos de los que parecen dar cuenta, ni las de aquellos otros que no rechazan la posibilidad de que se pudieran obtener medidas cuantitativas de los atributos psicológicos, pero que dudan de su significado y de su utilidad. Precisamente la disipación de estas dudas y la solución de estos problemas se abordan desde la teoría de la medición, pues ha sido en el ámbito más controvertido, en el de la Psicología, desde donde más esfuerzos se han hecho en este sentido en la segunda mitad del pasado siglo (Stevens, 1951; Suppes y Zinnes, 1963; Pfanzagl, 1968, 1971; Ellis, 1968; Krantz, Luce, Suppes y Tversky, 1971; Roberts, 1979), abriendo el debate y formalizando algunos de los conceptos básicos que se habían aceptado de forma natural y sin mayores exigencias en las denominadas ciencias físicas.

1.2. Clases y funciones

Las primeras nociones que es necesario comprender y dominar con agilidad son los conceptos ligados a la Medición y a la Teoría de la Medida. Estos conceptos son básicos y tan elementales como los de conjuntos y las operaciones que se pueden hacer con ellos.

Un conjunto, no es más que una colección arbitraria de elementos. A los elementos muchas veces se les denomina objetos, o bien, puntos. Se considera además que todos los conjuntos lo son de un espacio prefijado no vacío, al que denotaremos por la letra griega mayúscula Ω, para distinguirlo de sus elementos, que denotamos por la minúscula ω. Generalmente, todos los elementos del espacio Ω se denominan puntos y el conjunto de ellos se designa mediante $\{\omega\}$. Si los elementos ω son distinguibles, se les suele asignar un afijo, ya sea éste un apóstrofo (ω'), un subíndice (ω_i), un superíndice (ω^j), o ambos (ω_i^j) u otra notación cualquiera que se adopte para hacerlos distinguibles. A los conjuntos de puntos de Ω se les suele designar mediante letras mayúsculas: A, B, C, $\cdots$. Así, el conjunto Ω contiene a esos conjuntos y cuando un elemento particular pertenece a uno de ellos, por ejemplo al A, se dice que $\omega \in A$. Es posible que un elemento pertenezca a más de un conjunto. Un elemento ω puede pertenecer a A y a B, pero no pertenecer a C ($\omega \in A$; $\omega \in B$; $\omega \notin C$). Esto querría decir que los conjuntos A y B tienen al menos un elemento en común. Cuando dos conjuntos no tienen ningún elemento en común, esos conjuntos son "disjuntos". Si lo anteriormente dicho para un elemento, es cierto para todo elemento $\omega \in \Omega$, entonces el conjunto C sería disjunto con A y con B.

Conjuntos disjuntos darán intersecciones vacías. Es decir, al ser disjuntos no tendrán ningún elemento en común. Si, abusando del lenguaje, al conjunto que no contiene ningún elemento se le denomina conjunto vacío y se denota por $\varnothing$, entonces se puede decir que:

$A \cap C = \varnothing$ y que $B \cap C = \varnothing$

(Los símbolos $\cap$ y $\cup$ indican intersección y unión respectivamente).

En el caso en que el espacio Ω sólo tuviera los elementos de A, B y C, entonces se diría que:

$\Omega = A \cup B \cup C$

En cualquier caso, siempre se puede decir que el conjunto vacío está contenido en todos y cada uno de los conjuntos y se puede afirmar que:

$\varnothing \subset A \subset \Omega$

(Los símbolos $\subset$ y $\supset$ respectivamente indican "está contenido" y "contiene", y por $\not\subset$ no esta contenido)

Cuando todos los puntos de un conjunto A están contenidos en un conjunto B, entonces se dice que A es un subconjunto de B, y en tal caso se escribiría $A \subset B$, o bien $B \supset A$, o su equivalente: si $\forall\omega$, $\omega \in A$, $\omega \in B$, entonces $A \subset B$. Para cada conjunto A se dará la siguiente relación de inclusión citada: $\varnothing \subset A \subset \Omega$.

Estos conceptos elementales los introducimos, no porque no sean conocidos por el lector, sino a efectos didácticos, pues es preciso que el lector pase a definir, además de la unión y la intersección, la igualdad entre conjuntos, la diferencia A-B (conjunto que contiene todos los elementos de A que no pertenecen a B), manejando con soltura la simbología, por ejemplo:

i) Si se observa que las relaciones entre A y B están indicadas mediante los símbolos $A \subset B$ y $B \subset A$, entonces esto se identifica inmediatamente con que "todos los elementos de A están en B y viceversa", o sea, que $A = B$

ii) Si $\omega \in A$ y $\omega \notin B$, entonces es evidente que $\omega \in A\text{-}B$.

A la diferencia de Ω con cualquiera de sus subconjuntos, se le llama "conjunto complementario". Así Ω-A es el conjunto complementario de A y se designa por A^C.

Antes de continuar, el lector que no esté muy familiarizado en operar con conjuntos, debería de realizar algunos ejercicios, por ejemplo, realizar las operaciones que le lleven a comprobar que la relación de inclusión es reflexiva y transitiva, que la relación de igualdad es reflexiva, transitiva y simétrica y que las operaciones de unión y de intersección son asociativas, conmutativas y distributivas.

Se denomina *clase* a un conjunto particular de conjuntos de Ω y se les suele designar por A, B, C, ··· con o sin afijos.

La clase de todos los conjuntos de Ω se denomina espacio de conjuntos de Ω y se suele designar por $\mathcal{E}(\Omega)$ o bien por $S(\Omega)$. Una clase de conjuntos de Ω será un conjunto de $\mathcal{E}(\Omega)$ y todas las nociones y operaciones definidas en los conjuntos se aplican a las clases, consideradas ellas mismas como conjuntos. Las nociones de unión y de intersección, por tanto, se extienden de forma inmediata a clases arbitrarias, o sea, para conjuntos T no necesariamente de Ω. Ahora bien, si a cada $t \in T$ se le asigna un conjunto $A_t \subset \Omega$, la clase $\{A_t\}$ de todos esos conjuntos es una clase asignada al conjunto índice T.

1.3. Clases numerables

Se hace ya necesario introducir el concepto de *numerable* y pasar de las operaciones de unión y de intersección a la de adición y/o sustracción. Ello es debido a que las operaciones que se realizan sobre elementos de clases numerables tienen un papel relevante en el establecimiento de la noción de medida.

Un conjunto o una clase se dice *finito* cuando entre los elementos de ese conjunto o de esa clase y el conjunto de los n primeros números enteros y positivos $\{1, 2, \cdots, n\}$ se puede establecer una correspondencia biyectiva para algún valor de n y viceversa. Se dice que el conjunto es *enumerable*, cuando esa correspondencia puede establecerse con todos los enteros positivos, sin límite en su número, es decir, infinitos enteros positivos $\{1, 2, \cdots\}$. El conjunto, o la clase, se denomina *numerabl*e si es finito o enumerable.

1.4. Métrica y Espacios métricos

Los conceptos métricos se introducen en Ciencia para que, aprehendiendo el significado de algunas propiedades de los elementos y objetos, poder estudiarlos de forma sistematizada y, a ser posible, generalizada. La conexión entre el concepto *métrico* y la operación de *medir*, ha inducido a que en muchas ocasiones se haya afirmado, y así aparece en muchos textos, que "medir es asignar números a las cosas", o bien, que "medir es asignar números a las propiedades de los objetos". Estas definiciones, resultan incorrectas por imprecisas, pues la simple asignación de números no garantiza que esos números expresen las propiedades específicas de los elementos que se quieren expresar, es decir sus *magnitudes*, permitiendo además de su representación inequívoca, la manipulación experimental, así como la comparación con otras mediciones o magnitudes. Esto es, las representaciones numéricas de las propiedades de los objetos o elementos, a lo que se llama magnitudes, deben de ser resistentes a la manipulación experimental y ser susceptibles de operar matemáticamente con ellas, conservando tras esas operaciones su sentido y propiedades, permitiendo además que se puedan hacer con ellas comparaciones y predicciones.

En cuanto a los *espacios métricos*, escapa a los objetivos de esta obra tratar en extensión y profundidad los distintos tipos de espacios. Sin embargo, no es difícil aceptar, pues se puede comprender casi intuitivamente, que esas distinciones se basarán en las diferentes propiedades que posean unos y otros. Concretamente, esos conceptos están enraizados en los conceptos de elementos o puntos y la distancia que los relaciona. Por ejemplo, en el campo de los números reales $\mathbb{R} = (-\infty, +\infty)$, que es un conjunto medible en el que se pueden establecer los conceptos de límite, los distintos tipos de espacios a considerar están conectados con una métrica, que es la distancia euclídea que se puede establecer entre ellos. Entre dos puntos x, y de $\mathbb{R}$, se puede establecer una distancia, que se podría definir como $d = |y - x|$.

Si es E un conjunto, *una distancia en* E es una aplicación del producto cartesiano E x E en el conjunto $\mathbb{R}$ de los números reales, que tiene las propiedades siguientes:

(1) $d(x, y) \geq 0$ para cada par de elementos x, y de E
(2) la relación $d(x, y) = 0$ equivale a $x = y$
(3) $d(y, x) = d(x, y)$ para cada par de elementos de E
(4) $d(x, z) \leq d(x, y) + d(y, z)$ para cualesquiera tres elementos x, y, z de E. O sea, se da lo que se denomina desigualdad triangular.

De (4) por inducción se deduce que $d(x_1, x_n) \leq d(x_1, x_2) + d(x_2, x_3) + \cdots + d(x_{n-1}, x_n)$ para cada $n > 2$.

Si d es una distancia en E, entonces para tres elementos x, y, z de E, se cumple que

$$|d(x, z) - d(y, z)| \leq d(x, y)$$

Un *espacio métrico* es un conjunto E junto con una distancia dada en E.

Dos espacios métricos E y E´ se dicen *isométricos* cuando entre ellos existe una isometría. Si son d y d´ las distancias en E y E´, una biyección f de E sobre E´ se denomina *isometría* si para cada par de elementos de E

$$d'(f(x), f(y)) = d(x,y)$$

y la aplicación inversa f^{-1} es también una isometría de E´ sobre E.

La ventaja de tener espacios isométricos es que cada teorema que se demuestra que se cumple en E, que se refiera solo a distancias entre elementos de E, conduce de forma inmediata al correspondiente teorema en cada espacio isométrico E´, relacionando así las distancias de los elementos de E´, que son las imágenes obtenidas al aplicar f a los elementos de E.

Las distancias se definen de forma análoga en los espacios producto (se definen a continuación), sea cual fuere el número de ellos. Para dos espacios métricos E_1 y E_2, siendo d_1 y d_2 las respectivas distancias definidas en esos espacios E_1 y E_2, para cada par de puntos $x = (x_1, x_2)$, $y = (y_1, y_2)$ en $E = E_1 \times E_2$, sea

$$d(x, y) = \text{Máx}(d_1(x_1, y_1); d_2(x_2, y_2))$$

Se comprueba de forma inmediata que esta función satisface las propiedades anteriores, de la (1) a la (4), lo que indica que, en efecto, d(x, y) es una distancia en E. A este espacio métrico obtenido tomando la distancia d, se le denomina *espacio producto* de los dos espacios métricos originales E_1 y E_2.

La aplicación $(x_1, x_2) \rightarrow (x_2, x_1)$ de $E_1 \times E_2$ sobre $E_2 \times E_1$ es una isometría.

Son también distancias en E las funciones d´ y d´´ definidas mediante las siguientes relaciones:

$$d'(x, y) = d_1(x_1, y_1) + d_2(x_2, y_2)$$

$$d''(x, y) = \sqrt{(d_1(x_1, y_1))^2 + (d_2(x_2, y_2))^2}$$

Es por lo tanto equivalente tomar en E cualquiera de las distancias d, d´, d´´, para todas las cuestiones que se refieran a las propiedades topológicas de ese espacio.

La función $(x, y) \rightarrow |y - x|$ es una distancia en el conjunto de los números reales; el espacio métrico correspondiente se llama *recta real.* Cuando $\mathbb{R}$ se considera como un espacio métrico sin dar explícitamente la distancia, se entiende que se trata de la distancia antes definida.

En el espacio tridimensional $\mathbb{R}^3 = \mathbb{R} \times \mathbb{R} \times \mathbb{R}$, la *distancia euclídea* usual, para dos elementos $x = (x_1, x_2, x_3)$ e $y = (y_1, y_2, y_3)$ está definida por

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

y es trivial la verificación de los axiomas (1), (2) y (3) y el (4) se comprueba directamente mediante un simple cálculo.

La recta euclídea, (conjunto $\mathbb{R}$ en el que se ha definido d) puede poseer distintas propiedades, la continuidad uniforme, la monotonía y la complitud, entre otras (véase glosario de términos), y se tendrán distintos tipos de espacios dependiendo del número de esas propiedades que contengan. Los espacios pueden pasar de ser un espacio separable a ser un espacio métrico, a un espacio de Banach o a un espacio de Hilbert, en ese orden, a medida que crece el número de propiedades que se conservan.

Los elementos (puntos), conjuntos y clases que se consideren en el estudio, serán los del espacio en el que se esté representando el fenómeno y que, generalmente, será un espacio topológico[1].

1.5. Funciones medibles

El prototipo de todos los espacios, son los espacios construidos con números y, en consecuencia, las funciones cuyos valores son números, son el prototipo de todas las funciones.

Una función numérica X sobre un espacio Ω, es una función sobre Ω en $\overline{\mathbb{R}}$, definida asignando a cada elemento o punto $\omega \in \Omega$ un número único $x = X(\omega)$, que se debe leer como que "x es el valor de X en ω". Esa función X será finita, si se excluyen los valores infinitos.

Es aquí necesaria una reflexión: Con la notación $\overline{\mathbb{R}}$ hemos querido indicar que en $\mathbb{R}$ se incluyen valores infinitos, es decir, que la recta real $\mathbb{R} = (-\infty, +\infty)$ conteniendo sólo números finitos, se ha extendido a $\overline{\mathbb{R}} = \mathbb{R} = [-\infty, +\infty]$, lo que en algunos textos se denomina la *recta real ampliada*. Al representar el número x, generalmente entendemos que se refiere a un número real en el sentido usual, es decir, a un número finito, coincidiendo con nuestra representación mental inmediata.

[1] Al lector interesado lo remitimos a la obra traducida al español de M. Loève (1976) páginas 74-89, donde puede encontrar una exposición amplia de los conceptos asociados a espacios topológicos, así como de sus propiedades (la de Bolzano-Weierstrass, Heine-Borel, etc.)

Cuando se presenta el concepto de infinito, suele fallar la intuición y lo que se suele hacer, tanto en la vida ordinaria como en ciencia, es pasar de lo finito a lo infinito extrapolando, es decir, postulando que, lo que se observa en el caso finito, se seguirá cumpliendo en el caso infinito.

En la teoría de conjuntos, las uniones y las intersecciones de conjuntos A_n con n finito, se definen de forma obvia para el caso infinito:

Así, la unión $\bigcup_{n=1}^{\infty} A_n$ y la intersección $\bigcap_{n=1}^{\infty} A_n$ están definidas para n, donde n recorre un conjunto numerable de enteros.

Los números reales se definen en el rango:

$$+\infty \geq x \geq -\infty .$$

La necesidad de introducir números infinitos reside, entre otras razones, en que los límites de sucesiones monótonas existen siempre, pero esos límites pueden ser infinitos.

Si para una sucesión de números x_n existe el límite x cuando $n\to \infty$ (lo que se denota por $x = \lim x_n$, o bien por $x_n \to x$), se dice que x_n converge a x. Si ese valor x es infinito (bien sea $+\infty$, o sea $-\infty$), entonces se dice que la función diverge, bien a $+\infty$, o bien a $-\infty$.

La clase básica de conjuntos de $\mathbb{R}$ es la clase de ***intervalos***, de los que hay distintos tipos, según sean finitos, infinitos, abiertos, cerrados o semiabiertos, según se definan conteniendo valores finitos o no y conteniendo o no los extremos. Para dos valores cualesquiera a y b de $\mathbb{R}$, habrá cuatro tipos de intervalos finitos, éstos son:

(a, b) que es el conjunto de todos los puntos x tales que $a < x < b$
[a, b) que es el conjunto de todos los puntos x tales que $a \leq x < b$
(a, b] que es el conjunto de todos los puntos x tales que $a < x \leq b$
[a, b] que es el conjunto de todos los puntos x tales que $a \leq x \leq b$

El primer intervalo se dice que es abierto, los dos siguientes semiabiertos y el cuarto y último es un intervalo cerrado, pues contiene a ambos extremos.

Un par de conceptos que hay que tener claros y de los que se hace continuamente uso para definir las funciones medibles son los de *campo de Borel* y *conjuntos de Borel.*

El campo de Borel es el σ-campo minimal sobre la clase de todos los intervalos en $\mathbb{R}$ y sus elementos son los denominados conjuntos de Borel en $\mathbb{R}$.

Las operaciones numerables que se efectúen con elementos de una de las subclases indicadas anteriormente, dan cualquier elemento de las otras subclases. Por ejemplo, un intervalo abierto se puede obtener mediante la unión de intervalos semiabiertos a la derecha y un intervalo cerrado se puede obtener mediante la intersección de intervalos semiabiertos también a la derecha. Aquí el lector puede detenerse a comprobar, por ejemplo que $(4, 8) = [5, 8) \cup [6, 8)$ y que, en general, se puede escribir que $(a, b) = \cup\, [a + 1/n, b)$. Del mismo modo, el lector puede comprobar que el conjunto de todos los puntos x tales que $a \leq x \leq b$, o sea, que el intervalo cerrado $[a, b] = \cap\, [a, b + 1/n)$.

[2]De forma análoga se puede definir el campo de Borel sobre $\overline{\mathbb{R}}$ considerando cualquiera de los cuatro tipos de intervalos anteriores en los que $-\infty \leq a \leq b \leq +\infty$.

Los conceptos de σ-campo, medida y función medible, nacen tras los esfuerzos que se hacen en el siglo XIX y principios del XX, en el campo del análisis matemático, para ampliar el concepto de integración a clases cada vez más amplias de funciones.

Llegados a este punto es cuando probablemente se comprende mejor lo que es una función numérica, tal y como se ha definido al comienzo de este apartado y, salvo que se diga lo contrario, a todas las funciones medibles a las que se hace referencia, son funciones

[2] Las extensiones a los conceptos anteriores, especialmente su ampliación al espacio real N-dimensional, que se corresponde con un espacio producto de N rectas reales R, se puede ver en cualquier tratado de Teoría de la Medida y concretamente en Loève (1976).

numéricas, a las que se las suele representar mediante los símbolos X, Y, Z, etc.

Si, como se ha dicho, se cumple que $x = X(\omega)$, o sea, que es "x el valor de X en ω" entonces, para cada elemento o punto ω perteneciente a un conjunto A de Ω, sea ($\omega \in A$), $A \subset \Omega$, entonces:

a) que la función X es finita viene indicado por $|X| < \infty$;
b) que $X \geq 0$ sobre A, significa que $X(\omega) \geq 0$ para cada $\omega \in A$;
c) que $X_n \rightarrow X$ sobre A, lo que expresa es que $X_n(\omega) \rightarrow X(\omega)$ para cada $\omega \in A$, etc.

Por otro lado, debemos recordar lo que se conoce como *función inversa* de una función X. Si se tiene que $X \in S$, donde $S \subset \overline{\mathbb{R}}$ es el conjunto de todos $\omega \in \Omega$ para los cuales los valores $X(\omega)$ pertenecen al conjunto S, ese conjunto $[X \in S]$ es la *imagen inversa* de S y se designa como $X^{-1}(S)$. A la relación simbolizada por X^{-1} se le denomina *función inversa de* X, que es una aplicación de conjuntos en $\overline{\mathbb{R}}$ sobre conjuntos de Ω. Es evidente que las imágenes inversas de conjuntos disjuntos de $\overline{\mathbb{R}}$ son disjuntas y que las operaciones de X^{-1} con las de conjuntos, se pueden conmutar. Es decir, que la inversa de la unión o intersección de sucesos es igual a la unión o intersección del inverso, por ejemplo, $X^{-1}(\cup S_i) = \cup X^{-1}(S_i)$. También es fácil comprobar que la imagen inversa de un σ-campo es un σ-campo. Estas propiedades que son atribuibles a X cuando es unidimensional, se pueden extender a funciones $X = (X_1, X_2, \cdots, X_N)$ sobre Ω en un espacio real N-dimensional. En este espacio se pueden considerar funciones de funciones y al aplicar las funciones inversas de dos o más funciones sobre un subconjunto de S, las aplicaciones se hacen en orden inverso, es decir, operan de dentro a fuera, o si se quiere decir así, operan de derecha a izquierda. Esto es, para la función de función $gX(\omega) = g(X(\omega))$, la función inversa $(gX)^{-1}(S_i) = X^{-1}(g^{-1}(S_i)$, que simplificadamente se suele expresar como: $(gX)^{-1} = X^{-1}g^{-1}$

Uno de los mayores logros conseguidos en análisis matemático en los albores del siglo XX, que se debe fundamentalmente a autores como Borel y Lebesgue, es la introducción de una clase de funciones que es una *clase cerrada* con respecto a las operaciones usuales del análisis, como son las operaciones aritméticas, la formación de ínfimos, supremos y límites de sucesiones. Esto permite elegir en el dominio Ω de funciones un σ-campo $\mathcal{A}$ de conjuntos medibles. Al par (Ω , $\mathcal{A}$) se le denomina *espacio medible.* Así, si en el espacio final de funciones R se elige un σ-campo B, el campo de Borel en $\mathbb{R}$; el par ($\mathbb{R}$, $\mathcal{B}$) es una recta de Borel. Si se hubieran definido sobre $\overline{\mathbb{R}}$ y $\overline{\mathcal{B}}$, el par ($\overline{\mathbb{R}}$, $\overline{\mathcal{B}}$) es la recta de Borel ampliada. Por lo tanto, las funciones están definidas sobre un espacio medible (Ω , $\mathcal{A}$) en una línea de Borel, que puede ser ($\mathbb{R}$, $\mathcal{B}$) o ($\overline{\mathbb{R}}$, $\overline{\mathcal{B}}$), o bien ($\overline{\mathbb{R}}^N$, $\overline{\mathcal{B}}^N$) si el espacio es N-dimensional.

Se denominan *funciones medibles* a funciones tales que las imágenes inversas de todos los conjuntos de Borel sean conjuntos medibles. Si bien, basta exigir la medibilidad de las imágenes inversas de cualquier clase prefijada $\mathcal{C}$ (por ejemplo, la clase de todos los intervalos [-∞, x]) tal que el σ-campo minimal sobre $\mathcal{C}$ sea el campo de Borel. (Los conceptos de campo y σ-campo se pueden ver en el glosario que se incluye al final de esta obra).

Las nociones de *conjuntos medibles*, de *funciones medibles* y de *medida* se deben fundamentalmente a Lebesgue. Una amplia introducción sobre *medida* y *convergencias*, así como algunas de las propiedades de las funciones medibles y el teorema de medibilidad, lo puede consultar el lector interesado en profundizar en estos temas, entre otras obras, en la de Loève (1976, pp. 110-113 y 490-492).

2. La Medición

2.1 Metrización, estructuras, representación y escalas

La medición, como se ha dicho, aparece como consecuencia lógica de la metrización de los espacios, es decir, del establecimiento de las condiciones que hacen posible que se les asignen números a los objetos, como expresión de alguna de sus propiedades. Esto implica que se asume la suposición de que se puedan establecer condiciones de mensurabilidad. La medición se ha establecido, bien porque se quiere llegar a cuantificar y relacionar de algún modo las características más simples de los objetos (peso, longitud, color, ···), bien para hacer *universalmente* comparables esas características de los objetos, o bien, para contrastar teorías y establecer leyes, generando conocimiento científico, lo que en filosofía de la ciencia se ha llamado *ciencia normal* frente a otras formas de hacer ciencia (Kuhn, 1961, 1962; Santisteban, 2003). A través de la medición, del cálculo de errores, de la estimación de magnitudes, etc., se ha podido determinar en muchas ocasiones que una teoría establecida es falsa, como sucedió, por ejemplo, en astronomía donde hubo que desterrar las teorías pitagóricas mantenidas durante dos milenios.

En Psicología existe una época de gran polémica, que tiene lugar en los años 40 del pasado siglo, y que surge porque, tanto físicos, como filósofos y científicos de otros campos afines que se ocupan del problema de la medición, llegan a la conclusión, acertadamente, de que los modelos clásicos de medición no se muestran adecuados para la mayor parte de los fenómenos psicológicos. Lo que no fue correcto era concluir que, por lo tanto, la medición en Psicología, científicamente hablando, no era posible. Esto vino a rebatirlo en parte la teoría representacional, que desarrollan intensivamente, a partir de los años 50, psicólogos y otros científicos del ámbito de las ciencias sociales, que poseen los necesarios conocimientos de matemáticas y física para cambiar de forma sustancial los modelos clásicos de la medición física, para que se ajusten de forma adecuada a los propósitos de la Psicología científica.

Varios son los problemas estructurales que conlleva, al menos teóricamente, la teoría de la medición y, aunque a la medición como proceso experimental y de utilidad en el uso habitual no se le puede poner fecha de inicio, el desarrollo de sus fundamentos lógicos y axiomatización tiene un referente claro en Hölder (1901), si bien hay que considerar también como un buen precursor de este inicio de la moderna teoría de la medición, en lo que se llamó medición extensiva, la aportación de Helmholtz (1887). A raíz de la axiomática de Hölder (1901), muchos son los trabajos que se han ocupado del análisis lógico del proceso de medición, de la justificación de los diversos procedimientos y del significado y relevancia de sus resultados. Las teorías de Campbell (1920-1928) y la distinción que hace entre la medición fundamental y la derivada, lleva a que la medición fundamental comience con axiomas de tipo cualitativo sobre estructuras empíricas y entonces se pruebe la existencia de un teorema representacional en términos numéricos. Clásicamente, al menos desde el establecimiento de la teoría representacional[3], se han considerado en la medición al menos estos tres problemas: (1) el de la representación, (2) el de la unicidad y (3) el de la significación.

(1) *El problema de la representación*

Este problema surge ante la hipotética pregunta de si se pueden medir todos los atributos. Es decir, si las relaciones que se observan en el sistema empírico (en el mundo real) se pueden describir mediante sistemas formales, que es la definición de representación que dan Suppes y Zinnes (1963) mediante la relación homeomórfica establecida entre los dos sistemas relacionales. De una forma más general, podemos decir que el problema de la representación consiste en representar un sistema relacional empírico por un sistema relacional formal (el modelo), y si el modelo es numérico, la representación se llama medición.

Formalmente, si son A y B dos conjuntos de elementos en los que se han establecido las relaciones R y S respectivamente, se dice que un sistema $\alpha = \{A; R\}$ está representado por un sistema $\beta=\{B; S\}$ si existe una función f de A en B (cada elemento a de A tiene una imagen única f (a) en B) tal que para cualesquiera elementos a, b en A

$$a \; R \; b \Rightarrow f(a) \; S \; f(b)$$

O sea, que si el sistema α está representado por un sistema β existe una correspondencia f que aplica A en B, de modo que la relación R que se da entre elementos de A (sean a, b), implica que en B se de la relación S entre f(a) y f(b), que son las respectivas imágenes de a y b. Si la representación se da en ambos sentidos, es decir, si α

[3] La teoría representacional de la medición ha sido desarrollada por científicos de las ciencias del comportamiento y matemáticos fundamentalmente en la segunda mitad del sigo XX. Las bases de esta escuela de pensamiento se pueden resumir en la aceptación de dos de las condiciones de Campbell: la cuantificación basada en las propiedades empíricas de los objetos y que en ciencias sociales no se puede admitir con carácter general las operaciones de concatenación. Por lo tanto, rechazan que la única forma de cuantificación sea una operación de concatenación empírica, así como algunas de las propuestas de Stevens.

representa a β y a su vez β representa a α, entonces se dice que existe isomorfismo, o que los dos *sistemas son isomorfos*.

Si en el sistema empírico se establecen, por ejemplo, un conjunto de preferencias, y ese conjunto de preferencias se puede representar en el sistema formal, entonces se podrá construir una escala. Esto es, existe una escala de medida cuando existe un homomorfismo f del sistema relacional empírico α en el sistema relacional numérico β. Ahora bien, la relación empírica "preferido a" ($\succ$), conllevaría una representación numérica "mayor que" (>). Esto implica que la transitividad en la relación empírica es condición necesaria para poder representarla en el sistema numérico, concretamente por la relación "mayor que" en el sistema de los números reales.

Uno de los principales objetivos de la teoría representacional de la medición es el de investigar las condiciones, para establecer los pertinentes axiomas, bajo los cuales se pueden construir diversas representaciones numéricas. Un compendio de esos resultados se da mediante el llamado teorema de representación.

[4] Teorema: Sea A un conjunto finito y sea R una relación binaria conexa en A. Si R es transitiva, entonces, y solo entonces, existe una función f de A en los números reales tal que, para todo a y b de A es:

$$f(a) > f(b) \text{ si y solo si } a\,R\,b$$

Se suele formular el teorema de la representación en forma general indicando la existencia de una aplicación homomórfica entre un sistema relacional empírico y uno numérico.

(2) *El problema de la unicidad*

El teorema de la representación resuelve el problema de construcción de escalas, pero del teorema de representación no se deriva que haya una única escala, *los valores de la escala no están determinados por el modelo de medida*. Entonces habrá que preguntarse

[4] El enunciado y la demostración de este teorema se encuentran en los textos elementales tanto de medición como de Psicometría y/o construcción de escalas.

cómo se caracterizan las relaciones entre las diversas escalas numéricas de un teorema de representación, de cuanta libertad se dispone para la construcción de una escala que no de lugar a confusiones.

En primer lugar, se tienen las relaciones de preferencia a las que se les exige que sean transitivas, y además se tienen sus imágenes numéricas que satisfacen la relación "mayor que". Por ejemplo, supóngase que hay tres elementos con un orden estricto de preferencia y que satisfagan la desigualdad $f(a) > f(b) > f(c)$, entonces, tres números cualesquiera que satisfagan esa desigualdad formarían la escala, lo que indica que, en efecto, los valores de la escala no están determinados por el modelo de medida. Ahora bien, si cualquier conjunto de los valores de la escala satisface esa desigualdad, entonces se tiene una escala ordinal y se dice que la escala es única salvo una transformación que preserve el orden. Cualquier transformación de la escala que preserve el orden produce otra escala admisible. Así, dos escalas admisibles cualesquiera están relacionadas por una transformación que conserva el orden, esto es, por una transformación monótona.

Considérese ahora que además de establecerse un orden de preferencias estricto entre los elementos $a > b > c$ se pueden ordenar las diferencias de manera que haya la misma distancia entre elementos consecutivos, esto conllevaría que en el sistema numérico además de satisfacerse la desigualdad $f(a) > f(b) > f(c)$, también se cumple que $f(a) - f(b) = f(b) - f(c)$, que indica la formación de intervalos de igual tamaño. Ahora bien, es evidente que no toda transformación que preserva el orden cumple este requisito. O sea, que no toda transformación que preserve el orden es admisible. En este caso, para que la *transformación sea admisible*, no sólo debe de conservar el orden de los valores de la escala (en correspondencia con el de las preferencias), sino que también las diferencias y el orden de las diferencias entre esos valores. *Estas escalas son las denominadas escalas de intervalo y la única transformación que conserva la ordenación de los intervalos es una transformación lineal.* Bajo estas condiciones, una vez seleccionados dos valores de la escala, el resto de los valores están unívocamente determinados.

Por lo tanto, la definición formal de transformación admisible es necesaria para establecer el problema de la unicidad. Para los dos sistemas relacionales anteriormente definidos, α y β, para cualquier objeto $a \in A$ existe un único valor de escala f(a) en B, donde las relaciones en la escala reflejan las relaciones entre los objetos. Una transformación de la escala en B se dice admisible, si el sistema numérico que se genera en B al sustituir los valores de la escala original por los de la escala transformada, representa también el sistema empírico α.

Formalmente, para los dos sistemas $\alpha = \{A, R_1, R_2, \cdots, R_n\}$ y $\beta = \{B, S_1, S_2, \cdots, S_n\}$, donde B es el conjunto de los números reales, relacionados mediante la transformación f ya definida, una transformación g de B en sí mismo es una transformación admisible si para cualquier elemento $a \in A$, la transformación $f(a) = g(f(a))$ es también una representación de α por β.

El conjunto de todas las transformaciones admisibles en el sistema numérico, indica si se tiene o no unicidad de la medida y determina el tipo de escala.

(3) *El problema de la significación*

El problema de la significación aparece cuando se quieren interpretar correctamente las escalas, cuando se desea describir el significado de los números que representan las características de una determinada población, o bien cuando se trata de hacer inferencias basadas en los valores de esa escala. El problema es especialmente grave cuando se trata de justificar una asignación de números a objetos que no se derive de un teorema de representación establecido. El problema es especialmente difícil cuando los enunciados que incluyen valores numéricos no están soportados por un modelo explícito de medición, como es el caso de algunos atributos psicológicos.

La definición que dan Suppes y Zinnes (1963) es que un enunciado que incluye valores numéricos es formalmente significativo solamente si su verdad o falsedad permanece invariante bajo todas las transformaciones de los valores de la escala. Esto implica que se hace depender la significación de los modelos de medición concretos utilizados para obtener los valores numéricos.

Es este un problema difícil y creemos que poco desarrollado en la investigación psicológica. No obstante, creemos que el problema de la significación se resuelve, por un lado, construyendo escalas cuya interpretación inicial sea extremadamente clara y que las inferencias basadas en los valores de esa escala sean invariantes respecto a las transformaciones admisibles de la escala, y por otro, y en cualquier caso, la *significación* de los resultados depende de la *validez* de los supuestos subyacentes, aún en el caso en que los números que se hayan asignado a los objetos no estén soportados por un modelo de medición bien definido.

2.2. Medición directa e indirecta

Se llama medición directa a aquella que no depende de la medida de ninguna otra. El paradigma ya clásico de medición directa es la medida de la longitud. Sin embargo, la medición de otras magnitudes tales como la temperatura, la densidad o la velocidad, requiere la posibilidad de medir otras, para lograr la medición de la magnitud de la que esta en cuestión, es decir, sólo admiten mediciones indirectas. Estas mediciones, no solo deben cumplir algunos axiomas exigibles a la medida directa, como es la transitividad, sino que además conllevan la necesidad de que se definan ciertas leyes empíricas. Por ejemplo, la medida de la temperatura mediante un termómetro implica una medida directa de longitud (la altura que alcanza la columna de mercurio) y leyes para discriminar entre temperaturas.

La opinión de que hay ambigüedad en que una cantidad se pueda medir o no directamente no carece de sentido. En algún grado y en algún sentido hay una gran variedad de cantidades que se pueden medir directamente, aunque en la práctica científica esto se hace con muy pocas de ellas. La razón podemos ilustrarla con un ejemplo: es posible llegar a un consenso en que cierta persona es más inteligente o más brillante que otra, incluso se puede establecer la asimetría y la transitividad. Esto es, si la persona *a* es más brillante que la persona *b,* la *b* lo es menos que la *a*, y si la persona *a* es más brillante que la persona *b* y esta lo es más que la *c*, entonces la persona *a* es más brillante que la *c*. Pero en la práctica este juicio directo es extremadamente arriesgado y, en la práctica, no son estas las mediciones que se hacen de la inteligencia. La medida indirecta más común de la inteligencia es el cociente CI, así como otras inferidas a través de las puntuaciones en tests debidamente validados. En los albores del siglo XX, el físico Norman Campbell establece la distinción entre medición fundamental y derivada[5] (directa e indirecta), en relación con ellas dice que, aunque las unidades de algunas magnitudes fuese relativamente fácil medirlas directamente, sin embargo se prefiere definirlas en términos de medición indirecta (Campbell, 1928, 1957).

Creemos que el problema no radica en que la medición sea directa o indirecta, sino, por un lado, en la cuestión sustantiva de en qué grado la medida indirecta mide lo que se quiere medir, es decir, el problema de la *validez* de la medición y, por otro, en el grado de precisión de esas medidas.

La necesidad de distinguir que clase de información conllevan y representan las magnitudes, ha llevado a que se propongan distintas clasificaciones para designar los diferentes tipos de mediciones, según el procedimiento mediante el que se han llevado a cabo. Cuando para inferir números o leyes que representen unas propiedades hay que hacer uso de otras propiedades, a esa medida Campbell le ha llamado medida o *medición derivada*. Es el caso de la medida de la densidad, que se obtiene mediante la razón de la medida de la masa y la del volumen. Cuando los números se pueden asignar de acuerdo con las leyes naturales que representan propiedades de los objetos, sin necesidad de que ello presuponga la medición de ningunas otras variables, en el ámbito de la literatura en Psicología, autores como Campbell (1957) y Krantz et al. (1971) lo designan con el término de *medición* o medida *fundamenta*l.

---

[5] El lector interesado en una visión amplia de la teoría general de la medición derivada puede consultar la la obra de Suppes y Zinnes (1963); Sección 3.

Se puede considerar y definir la medida fundamental como la construcción de escalas estableciendo una relación isomórfica entre un sistema relacional empírico en un sistema relacional numérico. Por el contrario, en el sentido en que también usan el término Hempel (1952) y Suppes y Zinnes (1963), la medición derivada, genera una nueva escala a partir de otras escalas dadas. Hay autores como Pfanzagl (1968) que expresan sus dudas acerca de si es razonable considerar la medición derivada propiamente como medición, sugiriendo que el objetivo de la ciencia debe de ser construir escalas fundamentales para las propiedades de los objetos, con formulación de leyes empíricas y significado independientes, en lugar de tratar de satisfacer las propiedades de las escalas derivadas.

En la literatura también se considera una medición que se le suele denominar medición por "fiat", que responde sólo a las presunciones que se hacen sobre las relaciones que hay entre los objetos, junto al concepto que de ellos se tiene y generalmente esta medición se establece por definición. Ejemplos de ello tenemos en las medidas de la inteligencia, en sociología, etc. En esta medición la asignación de números está definida mediante alguna prescripción operacional, que no está basada ni sobre un homeomorfismo entre un sistema relacional empírico y uno numérico, ni sobre alguna relación funcional de escalas fundamentales, o sea, que tampoco es una medición derivada. Este tipo de medición se produce cuando un concepto precientífico, en un momento determinado en el desarrollo de esa ciencia, cobra una gran relevancia y no se tiene disponible al respecto, ningún procedimiento de escalamiento.

En la literatura psicológica en medición, se acuña un nuevo término con la aparición de la obra de Luce y Tukey en 1964, es el concepto de *medición conjunta*. Se llama *medición conjunta* a la forma de identificar la estructura aditiva de los atributos de forma indirecta, a través del establecimiento de relaciones de dos atributos con un tercero. Por ejemplo, es la forma en que se establecen las diferencias en densidad, observando las diferencias en volumen manteniendo la masa constante, revelándose así la estructura aditiva oculta en el atributo densidad.

Las estructuras de la medición conjunta son sistemas (X, Y, $\prec$), donde X e Y son conjuntos no vacíos y es $\prec$ un orden débil (por ejemplo, una relación transitiva y conectiva) sobre el producto cartesiano X x Y. Las representaciones aditivas consisten en dos funciones real-valuadas, por ejemplo las funciones f: $X \to \mathbb{R}$; g: $Y \to \mathbb{R}$ tales que los valores de ambas componentes se puedan sumar para reflejar el orden débil mediante la relación $\leq$ en los números reales $\mathbb{R}$.

Formalmente, para los pares (x,y) , (u,v) del espacio producto X x Y se tiene que:

$$(x,y) \prec (u,v) \quad \text{si y sólo si} \quad f(x) + g(y) \leq f(u) + g(v)$$

Las condiciones que garantizan esta relación resuelven el problema de la representación y la cuestión de la descripción de todas las posibles representaciones es el problema de la unicidad (Debreu, 1960; Luce y Tukey, 1964; Scott, 1964; Krantz, Luce, Suppes y Tversky, 1971, Narens, 1980, 1985; Suck, 1987) que, como hemos indicado anteriormente, es uno de los problemas más relevantes.

2.3. Escalas

En la teoría de la medición, una asignación consistente de los números reales a los elementos de una estructura empírica, se suele decir que es una escala para dicha estructura. Muchas de las estructuras empíricas que observamos en el mundo real, presentan un cierto orden natural y las escalas suelen reflejar ese orden, por lo que se les denomina escalas ordenadas. Dependiendo de la estructura que se mida y de como se hace la medición de esa estructura, se han considerado diferentes tipos de *escalas ordenadas*. Las más comunes, enumeradas desde la más débil a la más fuerte, en relación con las exigencias en las propiedades exigibles en la medición, son las conocidas como: escalas ordinales, de intervalo y de razón. No obstante, en la literatura se conocen distintas clasificaciones, muchas de ellas clásicas como la de Stevens (1951) que distingue entre escala nominal, ordinal, de intervalo y de razón. Es decir, añade a la dada aquí la denominada escala nominal, que hemos suprimido porque no es propiamente una escala. Las escalas nominales se forman, o bien denominando o nominando los diferentes objetos, o bien asignándoles un número para distinguirlos, o bien haciéndolo con las clases de objetos, agrupados según unas determinadas características. Aquí los numerales servirán para identificar las clases. Pero en nuestro concepto de medición, tal como entendemos y usamos este término, no ha lugar para las escalas nominales, pues los números que se asignan en las escalas se refieren al grado o a la cantidad relativa en que un objeto posee una cierta propiedad, y no a la propiedad del objeto en sí misma, como ocurre en las nominales.

Las escalas aparecen estrechamente ligadas a la medida, punto de vista que expresa Campbell (1928) cuando dice que la concepción de una magnitud es inseparable del orden de las características que posee. Otros autores, como Torgerson (1958), además de que consideran que ese orden debe de existir y estar definido, indican que la escala puede poseer un origen y las características de una distancia. Por ello, Torgerson propone que se distingan las escalas teniendo en cuenta el siguiente cuadro de doble entrada:

	Origen no natural	*Origen natural*
No distancia	Escala ordinal	Escala ordinal con origen natural
Distancia	Escala de intervalo	Escala de razón

O sea, que cualquier escala unidimensional será de uno de estos cuatro tipos de escalas: ordinal, ordinal con origen natural, de intervalo y de intervalo con origen natural (o de razón).

Una clasificación mas complicada es la que propone Coombs (Coombs, 1952; Coombs, Raiffa y Thrall, 1954), que añade a las de Stevens una quinta escala a la que denomina escala parcialmente ordenada. Bajo la concepción de Coombs, se debe pensar en las escalas considerando primero los objetos en sí mismos y después las distancias entre los objetos. Los objetos se pueden clasificar, ordenándolos, o bien parcialmente, o bien totalmente. Las distancias entre los objetos también pueden producir clasificaciones totales o parciales. Esto ha dado lugar a distinguir entre once diferentes clases de escalas. Cada una de estas escalas se nombra con dos términos, refiriéndose el primero al objeto y el segundo a la distancia, por ejemplo, escala nominal parcialmente ordenada. Nueve de las once escalas de Coombs caen dentro de la jerarquía de las escalas de intervalo de Stevens.

Una original e interesante interpretación de las escalas la proporciona Wright (1999) al conectarlas con la semiótica de Peirce, revisando y enriqueciendo la interpretación de Stevens (1946).

2.3.1. Escalas ordinales

Se denominan escalas ordinales a aquellas que tratan con estructuras en las que o bien sólo se conoce el orden de los elementos, o bien sólo el orden de los elementos es relevante. En este contexto, para un conjunto A de elementos, una estructura relacional $(A, \prec)$ se dice que es un orden débil (o un conjunto débilmente ordenado) si la relación $\prec$ es una relación binaria y transitiva sobre A. Para cualesquiera dos elementos diferentes $a, b \in A$, o es $a \prec b$, o bien es $b \prec a$, y a veces, por simplicidad, se dice que $(A, \prec)$ esta linealmente ordenado.

Si son x, y, z, tres magnitudes de una misma clase, los tres postulados que se suelen explicitar en este tipo de escalas son:

1. Si $x \neq y$, entonces, o es $x < y$, o es $y < x$.
2. Si es $x < y$, entonces es $x \neq y$
3. Si es $x < y$, siendo además $y < z$, entonces es $x < z$.

Si se considera el conjunto $\mathbb{R}$ de los números reales, una función $f :(A, \prec) \rightarrow (\mathbb{R}, \leq)$ es una escala ordinal (o simplemente una escala) de $(A, \prec)$ si y solo si $f(a) \leq f(b)$ en $\mathbb{R}$, siempre que para cualesquiera dos elementos $a,b \in A$, es $a \prec b$ en A. Y si existe una escala f de $(A, \prec)$, se dice que $(A, \prec)$ es *escalable*. Si A es un subconjunto de $\mathbb{R}$, que preserve el orden natural, entonces a la escala f de $(A, \prec)$ se dice que es una *función que preserva el orden* y existen funciones de A en $\mathbb{R}$ que son monótonas estrictamente decrecientes.

Escalas ordinales equivalentes : Dos escalas f y g de un orden débil arbitrario $(A, \prec)$, o sea dos funciones f, g : $(A, \prec) \rightarrow (\mathbb{R}, \leq)$, se dicen equivalentes cuando existe una función h que preserva el orden $h : f(A) \rightarrow \mathbb{R}$ tal que $g = h \circ f$, con lo que el dominio de la transformación que hace h depende de f. Esto es, si $x \in f(A)$, se elige $a \in A$ con $x = f(a)$, siendo $h(x) = g(a)$. Es fácil probar que definida así, la función h es una función que preserva el orden. Surge aquí el problema de la unicidad de las escalas, problema que apuntan, entre otros, Krantz et al. (1971) y diremos que el principal problema que se encuentra cuando se usan las escalas ordinales como sistemas de medición es precisamente la pobreza en las propiedades de unicidad (Narens, 1985). Por lo tanto, es necesario buscar condiciones más fuertes para definir la equivalencia entre escalas ordinales. Una de las soluciones que se han dado a este problema ha sido la de exigir el automorfismo. Concretamente, dos escalas f, g:$(A, \prec) \rightarrow (\mathbb{R}, \leq)$, se dicen equivalentes si existe un orden automórfico h de R tal que $g = h \circ f$. Esto lo que produce es que haya una relación de equivalencia sobre el conjunto de todas las escalas de $(A, \prec)$ y a partir de ahí, se pueden definir relaciones binarias, isomorfismos entre los retículos de ciertos subconjuntos de A (Droste, 1987a).

Por razones obvias, son las escalas que menos se han desarrollado desde el punto de vista de la teoría matemática, en comparación con las escalas de intervalo y/o de razón. Uno de los autores que, sin embargo se ha ocupado del papel que juegan estas escalas en el contexto de la teoría de la medición, ha sido Droste (1987a, b).

2.3.2. Escalas de intervalo y de razón

Los tipos de escala de mayor interés y utilidad son las de intervalo y las de razón. Son escalas únicas sobre un grupo φ de transformaciones lineales unidimensionales que son únicas sobre ciertas transformaciones lineales de $\mathbb{R}$. Esto es, si es φ el grupo de transformaciones lineales (positivas) de $\mathbb{R}$ en $\mathbb{R}$, se tiene una escala de intervalo o de razón cuando el conjunto de todas las transformaciones admisibles f es igual al conjunto de todas las aplicaciones de f(A) en $\mathbb{R}$ (φ f(A) $\rightarrow$ $\mathbb{R}$) que son de la forma φ: f (X) = a X + b, con $a \in \mathbb{R}^+$, $b \in \mathbb{R}$. La exigencia de que sea a > 0, lo que implica es que se tiene una relación de orden empírica que viene representada por el orden existente en $\mathbb{R}$, y si fuese a < 0 no se preserva el orden. En el caso en que sea b ≠ 0, se dice que la escala es de intervalo, y si b = 0, la escala es de razón.

Las escalas de intervalo y de razón implican la igualdad de intervalos, que se da si diferencias equivalentes entre las magnitudes representan la misma cantidad de diferencia en la propiedad que se mide. Las escalas de razón, adicionalmente admiten que la medida tenga cero absoluto, es decir, que el cero representa la ausencia de la propiedad que se mide.

El tipo de escala viene definido por el tipo de transformación admisible, esto es, por el tipo de transformaciones que mantienen su correcta representación. Téngase en cuenta que, por un lado, el nivel de escala de medida que se use, predetermina el modelo de escalamiento, y por otro, que el tipo de transformaciones admisibles de los valores de la escala que preservan la precisión en las predicciones de ese modelo, define el nivel de medida que se obtiene mediante la escala.

El desarrollo de la mayor parte de los modelos de escalamiento se ha dirigido a obtener escalas de intervalo y de razón. Un buen ejemplo en Psicometría es el modelo de escalamiento de Thurstone (1928), que hipotetiza como se puede medir un rasgo continuo mediante un test.

Un estudio detallado sobre la construcción de escalas de intervalo basadas sobre operaciones métricas, así como las basadas sobre distancias se pueden ver en Pfanzagl, 1968, capítulos 6 y 9, respectivamente.

2.3.3. Escalas derivadas de la respuesta

En los contextos experimentales, las escalas se derivan de las respuestas que dan los sujetos ante la presencia de un conjunto de estímulos. Las mayores diferencias entre los experimentos y, en consecuencia en la formalización de las correspondientes escalas, nacen de la consideración de donde radica el componente estocástico en el proceso, si sólo es atribuible al sujeto, o viene únicamente determinado por las características del estímulo, o por ambos.

Sea E un conjunto de estímulos, S un conjunto de sujetos y C el conjunto de todas las contestaciones o respuestas posibles a ese conjunto de estímulos.

Un experimento consiste en someter al sujeto $s_j \in S$ al estímulo $e_i \in E$ y registrar la respuesta $c_{ij} \in C$. La información sobre ese ensayo vendrá dada por la tripleta (s_j, e_i, c_{ij}). En general, presuponiendo que se pueden hacer repeticiones de ese experimento, con ensayos independientes, la descripción completa del experimento y de su resultado se puede representar en el espacio producto E x S x C, es decir, que esa descripción vendrá dada por la tripleta genérica (e,s,c) $\in$ E x S x C. Esto indica que para n repeticiones o ensayos independientes sucesivos (n = 1, 2,···), se pueden obtener una sucesión de pares (e_1, s_1); (e_1, s_2); ···; (e_2, s_2); ···; (e_i, s_j); ..., que se pueden tratar como variables independientes. Si se hacen esas consideraciones, a cada par (e_i, s_j) se le puede asociar una medida de probabilidad, de acuerdo con la distribución de probabilidades de las respuestas de los sujetos en ese experimento. Formalmente, se puede decir que a cada par (e, s) en el experimento, le corresponde una medida de probabilidad P(e, s, D), definida sobre el σ-Álgebra de subconjuntos $D \subset C$ que gobierna la distribución de los resultados de las realizaciones individuales en el experimento.

Si a un sujeto s en particular, sea el sujeto j, $s = s_j$, se le presentan un conjunto de estímulos $\{e_i\}$, a los que da respuesta, entonces ese conjunto de estímulos $\{e_i\}$ se puede ordenar según un criterio predefinido. Si ese mismo conjunto $\{e_i\}$ de estímulos se les presenta a distintos sujetos, se podrán obtener ordenaciones diferentes a partir de las respuestas de esos sujetos. Si el conjunto $\{e_i\}$ es de tamaño k, esto es, si el número de estímulos que se presentan es k, presuponiendo que el conjunto está formado por k estímulos diferentes, será k! el número un número máximo de ordenaciones diferentes que se podrá obtener. En un experimento concreto, cada una de esas ordenaciones aparecerá con una determinada frecuencia. En general, para un cierto número n de ensayos independientes, realizados sobre una población de N sujetos, se tendrá un número finito de puntos (conjunto de valores u ordenaciones) en C. La medida de probabilidad P(e, s, D) vendrá dada por un número finito de medidas de probabilidad, siendo $P(e_1, e_2, \cdots, e_k; s; c_1, c_2, \cdots, c_k)$ la probabilidad de que a los estímulos $\{e_1, e_2, \cdots, e_k\}$ les dé el sujeto $s = s_j$, la ordenación $\{c_1, c_2, \cdots, c_k\}$.

En el caso en que el elemento estocástico sea sólo el estímulo, es decir, que la respuesta del sujeto esté únicamente determinada por las características del estímulo, entonces se pueden considerar equivalentes tanto las muestras como las poblaciones, en cuanto a considerarlas homogéneas, y en la interpretación de los resultados, se pueden sustituir los términos muestra y/o población por el de sujeto. Esto conlleva también a que cuando se hable de diferentes sujetos se interprete que lo que se está considerando son diferentes poblaciones homogéneas.

A la probabilidad P(e, s, D), se le llama *medida de probabilidad de la respuesta*, y en el caso particular, que se ha expuesto anteriormente, en que el elemento aleatorio no es el sujeto, sino el estímulo, siendo P(e, s, D), independiente de s, entonces se suele hablar de *juicio*. Si la distribución de las respuestas no es independiente de s, se habla de *valuación*. Esto es, cuando existe un $e \in E$ y s_j, $s_g \in S$ tales que la igualdad en probabilidad $P(e, s_j, D) = P(e, s_g , D)$, no se da para todo D del σ-Álgebra , es cuando se habla de valuación.

Por otra parte, si la distribución de probabilidad de respuesta P(e, s, D) no está unívocamente determinada por el estímulo, sino que depende de las características del sujeto, por ejemplo, temperamento, se habla de *disposición* del sujeto, pues el estímulo se evalúa desde una determinada *disposición*. Esto tiene mayor interés en la construcción de escalas en las que en esa llamada disposición del sujeto intervienen valores éticos, morales, etc.

La distinción entre juicio y valuación se ha considerado muchas veces decisiva en la teoría de la medición, pues en el caso de los juicios, lo que se obtiene es una escala intersujeto para alguna propiedad de los estímulos. En el caso de la valuación lo que primeramente se obtiene es una escala para cada sujeto, pero sin embargo, se puede hacer que exista una escala común para los estímulos y las disposiciones de los sujetos tal que las respuestas del sujeto a los estímulos dependa de su posición relativa de pertenencia a esa escala común. Un ejemplo claro es cuando en una prueba de rendimiento se estiman simultáneamente los parámetros del sujeto y los de la prueba. A partir de los resultados experimentales, pues se puede obtener una escala común midiendo la dificultad de la tarea y la disposición de los sujetos, dependiendo de la posición del sujeto en relación con la tarea sobre esa escala. En definitiva, mientras que mediante juicios sólo se pueden conseguir mediciones intersujetos para las propiedades de los estímulos, la valuación puede conducir a la medición simultánea de los estímulos y de los sujetos.

Medición de estímulos: Si la probabilidad de la respuesta P(e, s, D), es independiente del sujeto $s \in S$, o sea, que la respuesta viene únicamente determinada por el estímulo, entonces es el estímulo el único elemento estocástico y se puede escribir simplemente que existe la probabilidad P(e, D) y si es posible una representación canónica, esto puede determinar una escala de intervalo para E.

Medición simultánea de estímulos y de sujetos: Cuando la probabilidad de la respuesta P(e, s, D), depende también de s, lo que se hace para construir la escala es definir una relación de equivalencia en el conjunto S de los sujetos, de manera que para cualesquiera dos elementos s_j y s_h de S, se considera que $s_j \approx s_h$ si y sólo si $P(e, s_j, D) = P(e, s_h, D)$ para todo $e \in E$.

Igualación de estímulos y pares comparados: En la literatura, las primeras aplicaciones de la teoría general que se ha desarrollado de la medición de estímulos y de estímulos y sujetos[6] ha sido la de la medición mediante igualación de estímulos y mediante comparación de pares. El primer método, pidiendo al sujeto que ajuste la variable estímulo a uno estándar dado, es un método esencialmente

[6] El desarrollo de estos apartados, con la enunciación de los teoremas y sus pertinentes demostraciones, pueden consultarse en la obra de J. Pfanzagl "Theory of Measurement", 1968. Cap. 11.

Fechneriano basado en el estudio de los valores medios de los errores. Los experimentos de Fechner (1860) generan distribuciones de probabilidad del tipo P(e,D) para diferentes valores del estímulo. A partir de las distribuciones de frecuencias construyó una escala en la que el rango intercuartílico era constante, lo que fue, para su época, un procedimiento no paramétrico excelente. La comparación entre pares de estímulos, se trata en varios apartados de esta obra, solamente señalar que cuando se pide a cualquier sujeto s ordenar dos estímulos dados e_i, e_k de acuerdo con algún criterio, no se admite la opción $e_i \approx e_k$, porque la indiferencia puede depender de ciertas características del sujeto que lo inhabilitan para discriminar entre ambos estímulos, por lo que en el procedimiento, se le fuerza a que admita la diferencia e indique si $e_i > e_k$, o bien si es $e_i < e_k$.

Para que exista medición y se pueda construir la escala, es necesario establecer formalmente las condiciones de monotonía, continuidad, etc., así como las transformaciones admisibles (Pfanzagl, 1968).

El problema de las comparaciones entre pares, se ha tratado en algunos trabajos añadiendo un supuesto adicional. Este supuesto adicional lo demandan, en general, los modelos de elección y han conllevando la inclusión de modelos logísticos. Esa suposición adicional se puede enunciar de la forma siguiente:

Si son e_i, e_k, e_h elementos de E, y es $P(e_i, e_k)$ la probabilidad que caracteriza la elección entre los pares (e_i, e_k), se supone que:

$$[P(e_i , e_k)/P(e_k , e_i)] \ . [P(e_k , e_h)/P(e_h, e_k)] = P(e_i , e_h) / P(e_h , e_i)$$

Este supuesto asume que $P(e_i,, e_k) > 0$ para todo $e_i,, e_k \in E$. Se supone además que no hay asimetría en los experimentos, es decir, que $P(e_i, e_k) + P\ (e_k, e_i) = 1$.

Esta condición adicional, se contempla en un gran varios trabajos de interés (Bradley y Terry, 1952; Bradley 1954a,b; 1955; Luce 1959), aún cuando es muy restrictiva y, en general, complica la construcción de la escala. Un caso particular de gran utilidad es cuando se usan los modelos logísticos. En la obra de Pfanzagl (1968) se dan los fundamentos de como se construye una escala m para los elementos e_i utilizando modelos logísticos: siendo e un elemento arbitrario $e \in E$. Esta escala está sujeta a transformaciones del tipo:

$$m(e_i)=\log\frac{P(e_i,e)}{P(e,e_i)}$$

donde

$$m'(e_i)=m(e_i)+\beta$$

cuando el elemento e cambia a un elemento e´.

$$\beta=\log\frac{P(e,e')}{P(e',e)}$$

En la misma obra se puede encontrar una discusión sobre las relaciones entre el método para escalar las *diferencias justamente perceptibles* y el método de comparación de pares. También se establece la generalización natural de la comparación entre pares cuando de acuerdo con algún criterio el sujeto elige uno (o más) estímulo(s) entre un conjunto $\{e_1, e_2, \cdots, e_k\}$ de $k > 1$ de ellos. Así, la elección entre pares es un caso particular con $k = 2$, porque la relación $e_1 \succ e_2$, se puede considerar como la elección de e_1 entre los elementos del conjunto $\{e_1, e_2\}$.

Respuestas dicotómicas: Consideremos aquí el caso en que el sujeto s, ante un estímulo e, está en una disposición que sólo permite dos tipos de respuesta, llamémosles positiva y negativa, asignándoles los valores uno y cero respectivamente. Esto es, las posibles respuestas pertenecen al conjunto $C=\{0, 1\}$ y la probabilidad de respuesta positiva será P(e, s), siendo e un elemento del conjunto E de los estímulos y s uno del conjunto S de los sujetos.

Para cada estímulo e (o ítem) se considera que es P(e, s) una función de la llamada “disposición”del sujeto, función a la que en el análisis de estructura latente se le ha llamado *traza* y en los tests mentales *curva característica del ítem*.

Ejemplos típicos los encontramos en las relaciones dosis-respuesta letal, en experimentación animal, donde la dosis es el estímulo, la "disposición" del sujeto es su resistencia, la propiedad del estímulo es la intensidad de la dosis y la respuesta es muerte o supervivencia. En el campo que nos ocupa, y puesto que nuestro interés está dirigido muy especialmente a la evaluación de la validez de los instrumentos de medida, podemos considerar como ejemplos la medición de las aptitudes, o de las actitudes. En estos casos, la "disposición" será respectivamente la aptitud, o la actitud del sujeto; el estímulo, la tarea o dar respuesta a determinadas cuestiones; la respuesta será, solución correcta u error, en la medición de aptitudes y, acuerdo o desacuerdo en la medición de actitudes. En el caso de un test de aptitud, la propiedad del estímulo será la dificultad.

El interés aquí está en que se puede demostrar que una representación canónica conduce simultáneamente a una escala para los estímulos y una escala para la disposición de los sujetos, que son únicas sobre una transformación lineal.

En primer lugar, hay que tener en cuenta que bajo el supuesto de respuesta dicotómica, se debe de distinguir entre el caso en que para todo s, P(e, s) es una función monótona de e, y cuando no lo es. Por ejemplo, la monotonía se da en el caso dosis-respuesta letal, pues una respuesta positiva para un estímulo e_i, implica que para otro estímulo e_h, tal que $e_i < e_h$, se producirá el mismo tipo de respuesta, o por el contrario, si para e_h hay supervivencia, entonces, salvo errores aleatorios, se dará para $e_i < e_h$. Lo mismo sucederá en los tests de aptitud, pues se supone que un sujeto capaz de resolver un problema o ejecutar una cierta tarea satisfactoriamente, también lo hará con otra de menor dificultad. A éstos estímulos los llamó Mosteller estímulos acumulativos, frente a los que llamó puntuales, como es el caso de las actitudes, donde no se da la monotonía. También Torgerson los llama puntuales, si bien, a los primeros los llama monótonos. Coombs, sin embargo, utiliza la terminología que es común fuera del ámbito de la Psicología, denominándoles monótono y no monótono.

Pues bien, si se considera que es m(e) una medida de la intensidad del estímulo y es n(s) una medida de la disposición del sujeto, para P(e, s) se puede dar una representación lineal (canónica) siempre que exista la función continua H : $\mathbb{R} \rightarrow \mathbb{R}$ y que los conjuntos m(e) y n(s) están conectados y que

$$P(e,s) = H\left(\frac{m(e) - n(s)}{\sigma(s)}\right)$$

Esta representación es única sobre todas las transformaciones

$$\begin{aligned} m^*(e) &= \alpha m(e) + \beta \\ n^*(s) &= \alpha n(s) + \gamma \end{aligned}$$

Una de las primeras aplicaciones prácticas de este modelo en medición del rendimiento y de la inteligencia, la realizó Thurstone (1926), suponiendo para H una distribución normal, si bien, este supuesto de normalidad, en general, no es necesario. Más aún, este supuesto, en muchos casos, no responde a la realidad.

En la determinación de P(e, s), como una función del estímulo, para un determinado sujeto s, las respuestas de los sujetos, y la respuesta del mismo sujeto a diferentes estímulos, se suponen que son independientes. Esto es lo que se ha generalizado en los modelos de la teoría de la respuesta al ítem (TRI) como la condición de *independencia local*, necesaria para que existan esos modelos. Solamente bajo esta suposición de independencia, la función de respuesta da una descripción completa de las reacciones del sujeto ante el estímulo, si bien, teóricamente, el caso de respuestas dependientes también se podría tratar mediante este método de representaciones canónicas.

En la práctica, si un sujeto se expone a k estímulos, y las respuesta a cada uno de ellos debe de ser dicotómica, el conjunto de las respuestas se puede representar mediante un vector v = $(v_1,..., v_k)$ compuesto sólo por unos y ceros. Este vector tiene 2^k posibles valores o configuraciones, denominados patrones de respuesta en la terminología del análisis de estructura latente. Si se pueden expresar sus probabilidades mediante la representación canónica:

$$P(e_1,\ldots,e_k,s,v_1,\ldots,v_k) = H_{v_1,\ldots,v_k}\left(m(e_1)-n(s),\ldots,m(e_k)-n(s)\right)$$

entonces las escalas m y n son únicas sobre transformaciones lineales.

La determinación de los valores de la escala $m(e_i)$, para todo estímulo e_i y para los diferentes sujetos, la realizó de forma práctica Rasch (1960), proponiendo un modelo en el que la estimación de los valores de las escalas se realiza utilizando estadísticos suficientes, esto es, cuando los valores a estimar los hace depender de las puntuaciones v_i solamente a través de

$$\sum_{i=1}^{k} v_i$$

que muestra ser un estadístico suficiente en la estimación de los valores de la escala.

2.3.4. Órdenes y medidas

Las relaciones asimétricas en el contexto de los juicios de similitud son una parte importante de la teoría de la atribución que han estudiado diversos autores, siendo el principal referente Tversky (Tversky, 1977; Tversky y Gati, 1978), quien establece una fuerte conexión entre la asimetría de una relación, lo saliente del estímulo y la noción de prototipo (sujeto y/o estímulo prototípico).

La importancia de revisar aquí la noción de asimetría, desde el punto de vista de los juicios de similitud, es porque en cierto tipo de contextos experimentales, las escalas numéricas que se establecen, se derivan de esos juicios y de esas relaciones. Por ejemplo, considérese que se perciben similitudes entre cuatro estímulos, sean: a, b, c, d, y que en un primer experimento se establece la similitud S entre a y b: S (a, b) y que en un segundo experimento se establece entre c y d: S(c, d). Se observa una *asimetría numérica* cuando $S(a,b) \neq S(b,a)$.

Supóngase además que los resultados del primer experimento conducen a la desigualdad $S(a,b) \leq S(c,d)$, o sea, que la similitud percibida entre los estímulos a y b es menor que la que se percibe entre los estímulos c y d. En ese caso, en relación con el segundo experimento se pueden dar una de las siguientes relaciones:

(1) $S(b,a) \leq S(d,c)$, o bien, es

(2) $S(d,c) \leq S(b,a)$

En el primer caso, la relación de asimetría que se había observado, permanece. La acción de la asimetría es puramente local, y no cambia en esencia lo percibido. En el segundo caso, se rompe la noción de similitud, vista ésta como un operador de orden que opera sobre los pares de estímulos.

Por lo tanto, en las tareas experimentales que impliquen juicios de similitud, hay que revisar las asimetrías numéricas resultantes y asegurarse de que se están utilizando las asimetrías sólo desde un punto de vista ordinal. En escalamiento y en medición, es así como tradicionalmente se contemplan, implícita o explícitamente, dado que los análisis de similitud a menudo se fundamentan en órdenes de similitud sobre pares.

Una justificación formalizada de estas ideas, implementada con resultados experimentales, se puede ver en Barthélémy (1984).

Un tema fundamental en el establecimiento de las relaciones de orden es *la modelización de las preferencias.* Para ello, se hace uso de los conceptos de preorden, semiórdenes, intervalos ordenados, pseudo-ordenes y ordenes difusos y se suele asumir que las relaciones son completas, es decir, que son comparables. Sin embargo, en la práctica, aparecen muchos casos en que esas propiedades no se cumplen. Situaciones en que los elementos no son comparables aparecen de forma natural, por ejemplo, en múltiples problemas de toma de decisiones, y dentro de éstas, especialmente, en problemas multicriterio. Son relativamente corriente los casos en que, al definir relaciones de orden parcial en intervalos, solamente se puedan encontrar relaciones de preferencia, de indiferencia y de no comparabilidad, por lo que es obligado redefinir las relaciones de manera que sean compatibles con una representación numérica mediante propiedades que sean fáciles de verificar.

Para dar las representaciones numéricas de preórdenes totales y parciales consideremos lo siguiente:

Sea P una relación asimétrica y sea I una relación reflexiva y simétrica sobre un conjunto finito A. La pareja (P,I) es un *preorden completo* (o total), si y sólo si:

a) $\forall a,b \in A$; se cumple una de las tres relaciones: aPb, o bien, bPa, o bien, aIb
b) P es transitiva
c) I es transitiva

Ese preorden completo tendrá representación numérica si y sólo si existe una función g tal que, $\forall a,b \in A$, se cumplan las siguientes relaciones

$a \; P \; b \Leftrightarrow g(a) > g(b)$
$a \; I \; b \Leftrightarrow g(a) = g(b)$

La pareja (P,I) es un *preorden parcial* con representación numérica, si y sólo existe una función g tal que, $\forall a,b \in A$, se cumpla que

$a \; P \; b \Rightarrow g(a) > g(b)$
$a \; I \; b \Leftrightarrow g(a) = g(b)$

Por lo tanto, el concepto de preorden parcial es suficiente, pero no necesario para tener al menos una representación numérica, porque esta representación no implica la transitividad de P. De hecho, la condición necesaria y suficiente para tener esta representación numérica es que la relación I sea transitiva y que la relación P no tenga ningún circuito. (Existe un circuito cuando existen elementos A, sean $a_1, a_2, \cdots, a_n$ tales que para cualquier i existe la relación $a_{i-1} \; P \; a_i$, y también existe la relación $a_n \; P \; a_1$).

Una generalización de los órdenes de intervalos la da Doignon (Doignon, 1984; Doignon, Ducamp y Falmagne, 1984) para modelizar las preferencias individuales, considerando una relación P de asimetría entre los elementos de un conjunto A de objetos. La relación P contiene los pares ordenados (a, b) de manera que es el objeto b estrictamente preferido al objeto a. También se define sobre A la relación de indiferencia I que se supone que es reflexiva, simétrica y disjunta con P. La generalización se hace en tres direcciones: por un lado, se generaliza el clásico modelo de umbral de Luce (1956), a múltiples umbrales, por otro, se generalizan las estructuras parciales, considerando el par de relaciones (P,I) de preferencias e indiferencias sobre A, siendo la relación P asimétrica y la relación I simétrica y reflexiva, pero sin suponer que esas relaciones se puedan establecer entre cualquier par de objetos. Esta es la situación, como se ha dicho antes, de que un individuo puede considerar que dos objetos no son comparables, o que no tiene la información o los elementos de juicio suficientes para establecer su comparabilidad. Esto es lo que conduce a Roy (1980) a considerar representaciones numéricas con las dobles implicaciones señaladas anteriormente, que reemplazan a las implicaciones en una sólo dirección. Por último, se generaliza al caso infinito, es decir, los resultados de (1) y (2) se consideran generalizables al caso de infinitos conjuntos A, bajo ciertas restricciones. Todo ello se hace basándose en unas relaciones para las que los autores (Doignon, Ducamp y Falmagne, 1983) acuñan un nuevo nombre, llamándolas *biórdenes* (relaciones de Ferrers para otros autores, Monjardet, 1978).

2.3.5. Transformaciones admisibles

Independientemente de la clasificación que se adopte de las escalas, una cuestión fundamental a tener en cuenta es la de la invarianza de las características de los objetos que la escala representa cuando se hacen las transformaciones algebraicas. Es decir, hay que tener en cuenta las transformaciones admisibles para cada tipo de escala. Tomando en consideración como se han definido las diferentes escalas y sus propiedades, es fácil deducir las transformaciones admisibles para cada una de ellas, que son:

a) Escala ordinal: transformaciones monótonas
b) Escala de intervalo: transformaciones lineales de la forma $g(x) = a\ f(x) + b$
c) Escalas de razón: transformaciones de la forma $g(x) = a\ f(x)$ donde x representa un elemento genérico del sistema empírico, cuya imagen en el sistema numérico es $f(x)$.

Un ejemplo sencillo de algunas de estas transformaciones se tiene cuando se pone en relación el área de un círculo con su radio. Si se considera una escala ordinal, esta relación es monótona, siendo el área una función monótona del radio, pues considerando entre ellos sólo una relación ordinal, cuando el radio crece, el área crece. Ahora bien, al ser el área del circulo una función cuadrática del radio, se puede considerar una escala de razón, pues se mantiene la proporcionalidad entre ambas magnitudes con cambios cualesquiera de la longitud del radio, siendo π (el número pi) la constante de proporcionalidad.

Un ejemplo de escala de intervalo se obtiene si se pone en relación el número de palabras correctamente leídas con la capacidad lectora. La escala de puntuaciones es una escala de intervalo si la diferencia de X puntos en esa escala representa la misma cantidad de diferencia en la capacidad lectora. En ese caso, las relaciones se conservarán con transformaciones lineales.

Si las transformaciones admisibles para una escala son el producto por una constante, la representación será una *escala proporcional.* Si las transformaciones admisibles son las lineales del tipo $g(x) = a\ f(x) + b$, la representación es una *escala de intervalos*, y si la transformación lineal es del tipo $g(x) = f(x) + b$, se tendrá una *escala de intervalos absolutos*, en el sentido de que permanece constante la diferencia entre los valores.

Si las transformaciones admisibles son las transformaciones exponenciales del tipo $g(x) = a(f(x))^n$, entonces se generarán *escalas logarítmicas de intervalos,* ya que la relación entre ellas es lineal, pues $\log g(x) = \log a + n \log f(x)$. Para transformaciones exponenciales $g(x) = (f(x))^n$, se generarán *escalas logarítmicas de proporciones,* caracterizadas por permanecer constante el cociente de los logaritmos de los valores, pues es $n = (\log g(x) / \log f(x))$.

MEDICIÓN Y VALIDEZ: MODELOS PSICOMÉTRICOS AVANZADOS

1. Introducción

En el capítulo precedente se han abordado los conceptos básicos de la teoría de la medida y posibles soluciones para la medición de constructos como la medida aditiva conjunta que, como veremos en este capítulo, constituye uno de los acercamientos más prometedores para la medición psicológica y, en concreto, para la medición mediante tests o instrumentos psicométricos.

En este capítulo se trata de conectar los temas de validez con los de la medida, estudiando aquellas aproximaciones psicométricas que permiten realizar una medición rigurosa, fundamentada y válida. El problema de la medida y el de la validez generalmente se han estudiado de modo separado, como si el establecer el significado de las puntuaciones de un determinado test nada o poco tuviera que ver con la cuestión de si es correcto o no el procedimiento de medida. Los estudios acerca de la validez se han preocupado más por buscar la justificación en el uso de los instrumentos, que de averiguar si la medición era o no correcta, en tanto que la teoría de la medida sólo se ha ocupado de establecer los principios de lo mensurable y de los procedimientos legítimos para poder medir.

Ya en la primera mitad del pasado siglo, los psicómetras se dieron cuenta de que muchos de los tests psicológicos estaban construidos siguiendo criterios subjetivos y, generalmente, sin un claro apoyo teórico. Por lo tanto, la calidad de esos tests dependía de la intuición de sus creadores, que difícilmente se podrían ajustar a los exigentes criterios de la medición científica expuestos por Thurstone entre 1925 y 1932, y que han sido recogidas por Wright (1999), estos son :

1. *Unidimensionalidad*: Una característica universal de toda medida, ya sea de un objeto o de una entidad, es que represente a un solo atributo del objeto medido (Thurstone 1931).
2. *Linealidad*: La idea de medida implica la existencia o construcción de un continuo lineal en el que se puedan ordenar las cantidades (Thurstone y Chave, 1929). Las puntuaciones directas de los tests, ya sean de formato dicotómico, créditos parciales, escalas de evaluación o escalas Likert, raramente constituyen escalas lineales, puesto que suelen estar sesgadas a favor de los valores centrales y contra los valores extremos. Por lo tanto, todo procedimiento estadístico lineal (ANOVA, regresión, análisis factorial, ecuaciones estructurales, etc) que utilice como entrada puntuaciones directas conducirá probablemente a resultados equivocados y dependientes de la muestra y, por lo tanto, la inferencia será ambigua. Parafraseando a Wright (1999): *para aplicar métodos estadísticos lineales, antes debe de aplicarse un modelo que permita obtener medidas lineales a partir de los datos observados.*
3. *Abstracción*: El continuo lineal implícito en toda medida siempre es una abstracción. Es una falacia pensar que cualquier cosa puede constituir una unidad de medida (por ejemplo, un trozo de madera o una vara). Para el establecimiento de unidades de medida previamente habrá que realizar un proceso de abstracción (Thurstone, 1931).
4. *Invarianza*: La unicidad de medida supone que sea replicable e intercambiable en las diferentes partes del continuo (Thurstone, 1931).

5. *Calibración de los ítems "libre de muestra" (sample-free)*: La estimación de las características de los ítems debe ser independiente de las características concretas de las muestras de sujetos. Es decir, el instrumento de medida no debería verse seriamente afectado por el objeto medido, la escala debe trascender al grupo que constituye la muestra normativa o de calibración (Thurstone, 1928).
6. *Estimación de la aptitud "libre de test" (test-free)*: El nivel estimado del sujeto en la escala no debe de depender de los ítems concretos que constituyen el test. Si se omiten algunos ítems, o se realiza sólo parte de la prueba, o se ordenan de otra manera los ítems etc, estas manipulaciones no deben de afectar a la puntuación estimada de la persona en la escala de aptitud (Thurstone 1926).

Alcanzar los requisitos exigidos por Thurstone (1926) se antojó inalcanzable para la mayoría de los psicómetras (Guttman 1950, supuso una pequeña excepción), iniciándose una huída hacia delante, con el convencimiento de que los tests que usan los psicólogos pueden ser útiles (válidos) pero difícilmente podrán cumplir los rígidos criterios que exige la medida científica. Michell (1999) hace un recorrido pormenorizado del debate de la medida en Psicología censurando a Stevens (1939; 1946) por lo inconcreto de su afirmación de que medir consiste simplemente *en asignar valores numéricos a los objetos conforme a una regla*. Sin embargo, la cuestión de la fundamentación métrica, como hemos analizado en capítulos anteriores, no tiene nada de trivial y la medida solo estará justificada cuando la asignación (homomorfismo) esté sustentada en transformaciones admisibles. El propio Stevens, tras establecer los cuatro tipos de escalas (nominal, ordinal, de intervalo y de razón) se dio cuenta de esta dificultad al reconocer que los tests, a lo sumo, pueden alcanzar un nivel de medida ordinal. Lord resolvió, momentáneamente, el problema de la tratabilidad matemática de los datos obtenidos con los tests, sean o no ordinales, con su curioso argumento de que *los números no saben de donde vienen* (ver Santisteban y Alvarado, 2001, capítulo 1).

Con estos antecedentes, hasta en los más prestigiosos manuales de Psicometría (e.g. Lord y Novik, 1968) se ha obviado el debate, considerando que las medidas que se obtienen al aplicar los tests, o al menos aquellos que muestran propiedades psicométricas aceptables, se pueden tratar como medidas en una escala de intervalo.

En contraste con este modo de proceder, revisaremos modelos psicométricos que se han preocupado de la cuestión de la fundamentanción métrica, que han demostrado que es posible medir, e inferir explicaciones para los constructos psicológicos, de una forma rigurosa y exigente, aunando validación del constructo y calidad métrica. Antes de pasar a revisar estos modelos, repasemos brevemente los criterios formales que se han ido estableciendo desde la Psicología matemática para la medida cuantitativa y que permitirían obtener medidas con las características descritas por Thurstone:

a) *Concatenación*: El principio (o fundamento) en el que descansan las medidas físicas es la concatenación (Campbell, 1920). Puesto que a la medida de atributos psicológicos no se puede aplicar este principio, fueron necesarios nuevos avances en la teoría de la medida para que también la medida psicométrica pudiera alcanzar los requisitos de la medición científica.

b) *La suficiencia*: Fisher (1920) desarrolla el procedimiento de estimación maximoversimil que permitirá la obtención de estadísticos suficientes. El trabajo de Fisher tiene una importante implicación para la construcción de medidas cuantitativas, puesto que, la posibilidad de reemplazar el parámetro por su estimador suficiente permite construir medidas generalizables: cuando un modelo carece de estadísticos suficientes para sus parámetros, se está ante un modelo de escasa calidad métrica, ya que es incapaz de hacer estimaciones independientemente para los parámetros.

c) *Divisibilidad*: Paul Levy (1937) demostró que la construcción mediante variables aleatorias de continuos *estables* para determinar que lo *numerable* requiere de parámetros infinitamente divisibles. Wright (1999) observa que la divisibilidad de Levy es logarítmicamente equivalente a la aditividad conjunta (Luce y Tukey 1964) y que la divisibilidad no sólo está implícita en la suficiencia ,como mostró Kolmogorov (1950), sino que además sería la condición necesaria y suficiente para construir cualquier espacio cuantitativo (Bookstein, 1992, 1996).

d) *Aditividad*: Luce y Tukey (1964) en su formulación de la aditividad conjunta dan una respuesta a la exigencia de concatenación de Campbell para la medida psicológica. Al respecto dicen que "*cuando no existen operaciones naturales de concatenación, se debe de intentar descubrir otras maneras de medir factores y respuestas de forma que los "efectos" de los diferentes factores sean aditivos"*. También afirman que "*existe una estrecha relación entre la medida conjunta y el establecimiento de medidas de respuesta en una tabla de doble entrada... cuando "los efectos de las columnas" y "los efectos de las filas" son aditivos. De hecho cuando se descubren este tipo de medidas....se puede entender como el descubrimiento, vía medida conjunta, de medidas fundamentales"* (Luce y Tukey 1964).

2. Propiedades métricas del modelo de Rasch

El modelo de Rasch (1960) es un ejemplo de medida aditiva conjunta con una estructura estocástica subyacente, por lo que pueden aplicarse los procedimientos generales de inferencia estadística para poner a prueba la bondad de ajuste del modelo a los datos empíricos. El modelo de Rasch proporciona la posibilidad de demostrar la objetividad de la medida, la "objetividad específica", como característica del modelo que permite la comparación de los sujetos, con independencia de los instrumentos (estímulos) utilizados para medirlos, así como la comparación entre instrumentos, con independencia de los sujetos a los que se aplicó. Por lo tanto, el modelo de Rasch es un ejemplo concreto de cómo la medida aditiva conjunta puede aplicarse a datos empíricos.

La teoría de la medida conjunta nace motivada por la evidencia de que, como se ha dicho anteriormente, los clásicos modelos generados para la medida física, no eran adecuados para la medición de la mayor parte de los fenómenos psicológicos. Uno de los mayores escollos era el de la concatenación, pues aunque no todas las propiedades físicas la admiten (por ejemplo, la temperatura o la velocidad), es mucho más general en los fenómenos físicos que en los psicológicos. La dificultad en aplicar la concatenación a los atributos psicológicos fue la causa de que inicialmente se pensara (e.g. Stevens, 1939) que difícilmente se podría superar el nivel ordinal de medida. Sin embargo, posteriores investigaciones en teoría de la medida, han llevado a una conclusión diferente, puesto que ahora sabemos que la operación de concatenación empírica no es necesaria para la medida en escala de intervalo, proponiéndose distintos modelos que permiten obtener escalas de intervalo (Coombs, Dawes, y Tversky, 1970), entre los que destaca por su interés psicométrico la medida aditiva conjunta. Esto vino a contradecir una de las condiciones de Campbell, que la única forma de cuantificación es una operación de concatenación empírica.

La medida aditiva conjunta establece la manera de ordenar una variable dependiente que varía con el efecto conjunto de dos o más variables independientes. Por ejemplo, para un primer análisis de la aditividad conjunta de dos variable, puede realizarse un análisis de varianza, para poner a prueba si la variable dependiente se puede representar como la suma de los efectos de filas y columnas. No obstante, si se intenta obtener este resultado utilizando como variable dependiente la puntuación directa del test, probablemente no se obtendrá el resultado deseado, siendo necesario realizar una transformación (por ejemplo, mediante el modelo de Rasch) o re-escalamiento monótono de la variable dependiente (variable ordinal) que elimine la interacción entre las variables independientes (Fischer, 1995a; 1995b). Cuando los datos se ajustan a un modelo de Rasch, el resultado es que la variable dependiente, una vez transformada, y las variables independientes (parámetros de las personas y de los ítems) estarán conjuntamente (de aquí, el término conjunta) representadas en una escala de intervalo, siendo la unidad de medida común a personas e ítems.

Para considerar el modelo de Rasch como un ejemplo de modelo de aditividad conjunta, considérese, por ejemplo, que se tiene una muestra de individuos que realizan una prueba de inteligencia.

Sea P_{ij}, la probabilidad de que la persona *i* responda correctamente al ítem *j,* asumiendo que esa probabilidad depende únicamente de dos factores: de la aptitud de la persona *i* (θ_i) y de la dificultad del ítem *j* (β_j). Además, para poder estimar, a partir de los datos, los parámetros aptitud y dificultad, será necesario recurrir a algún modelo matemático que permita realizar una transformación monótona, M, de modo que se obtenga un modelo aditivo: $M(P_{ij})= \theta_i + \beta_j$

Cuando se muestra la bondad de ajuste de los datos al modelo aditivo, estamos ante una representación aditiva de las probabilidades P_{ij} y ,en principio, todo modelo o transformación de los datos en probabilidades será válido siempre que el resultado sea una representación aditiva.

En la teoría psicométrica, hace ya casi medio siglo que se propusieron dos modelos de la TRI unidimensionales que permiten obtener este resultado: el modelo de ojiva normal y el modelo de Rasch (Santisteban y Alvarado, 2001). La representación aditiva $M(P_{ij})= \theta_i + \beta_j$ se corresponde con la aplicación del modelo de ojiva normal, mientras que el uso del modelo de Rasch conduce a un modelo $M(P_{ij})= \ln[P_{ij}/(1-P_{ij})]$ denominado *logit* que se corresponde con una representación $M(P_{ij})= \theta_i - \beta_j$.

En consecuencia, tanto el modelo de ojiva normal como el de Rasch, permiten obtener un modelo aditivo de los parámetros aptitud y dificultad, además, el modelo de Rasch tiene propiedades estadísticas y métricas muy deseables, como la suficiencia o la objetividad específica que se analizarán a continuación.

Estudiaremos las propiedades métricas del modelo de Rasch[7] (RM) siguiendo el esquema que se muestra en la Figura 1. En este esquema se concretan los elementos a investigar para conocer las propiedades de las escalas en los modelos psicométricos. Recordemos que en la teoría axiomática de la medida (ver capítulo 1), una escala se genera con el establecimiento (consiste en establecer) de un homomorfismo entre el sistema relacional empírico y otro numérico, mediante una transformación admisible, Φ, que garantice la *unicidad*

---

[7] A partir de aquí utilizaremos para referirnos al modelo de Rasch el acrónimo RM (Rasch Model).

de la escala. De modo que la medida, de intervalo, exige que la representación de los objetos empíricos en números reales admita la transformación: $x \rightarrow ax + b$, para $a > 0$ y $b \in \mathbb{R}$.

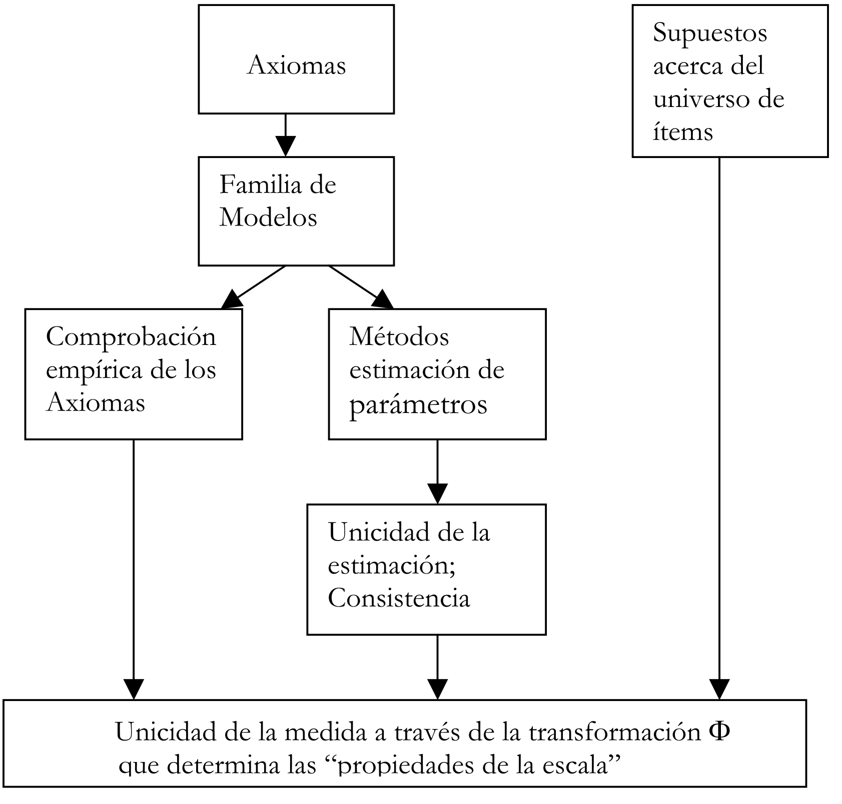


FIGURA 1. Elementos para una investigación de las propiedades en las escalas TRI (tomado de Fischer, 1995a).

Las propiedades de los modelos de la familia RM se derivan de un conjunto de supuestos, de los que los tres primeros axiomas son los supuestos fundamentales o "técnicos", en palabras de Fischer (1995a), y estos son: (i) continuidad y monotonía de las CCIs, (ii) las CCIs son funciones de probabilidad e (iii) independencia local. A estos tres supuestos se adicionan otros supuestos convenientes para la demostración de las distintas propiedades del RM.

En la Figura 1 se indica que la estimación de parámetros es el procedimiento que se sigue para hacer transformaciones admisibles de los objetos empíricos en valores numéricos, habiéndose desarrollado métodos específicos para los RM: máxima verosimilitud condicional (CML). Estos métodos han mostrado una gran consistencia bajo distintas condiciones de la distribución de los parámetros personales (o aptitud). Además, se han desarrollado un buen número de paquetes informáticos para la implementación del modelo como RASCAL (Assessment Systems Corporation, 1995) o WINSTEPS (Wright y Linacre, 1998). Puesto que existen métodos e implementaciones informáticas que permiten hacer estimaciones de los parámetros con gran exactitud, nos centraremos ahora en mostrar como de unos pocos axiomas pueden derivarse las propiedades métricas de los modelos de Rasch.

2.1 Suficiencia

En el modelo de Rasch el número de aciertos de un sujeto v o puntuación directa R_v obtenida en un conjunto de ítems i con parámetros β_i, es un estimador suficiente del parámetro aptitud θ_v, como ya se indicó para un determinado conjunto de ítems con parámetros β_i. Como se puede colegir del apartado 2.3.3. del capítulo anterior (Respuesta dicotómica), esta propiedad es muy atractiva y supone que el estadístico R_v permite *medir* θ_v , sin necesidad de hacer referencia a los parámetros de los ítems β_i concretos utilizados para su estimación. Una vez establecido el estimador suficiente de la aptitud, para un determinado patrón de respuestas $\boldsymbol{x}_v$, éste ya no contiene información adicional acerca de θ_v más allá de $R_v=\Sigma_i X_{vi}$.

Ahora bien, ¿qué modelos psicométricos son compatibles con la propiedad de suficiencia?. Demostraremos a continuación que de la propiedad de suficiencia es posible derivar una importante familia de modelos: los modelos logísticos tipo Rasch.

La demostración parte de la existencia de un determinado conjunto de ítems I_i que se ajustan a un modelo unidimensional de la TRI en el que se satisfacen los siguientes cuatro axiomas fundamentales:

Axioma (i). Continuidad y monotonía: las Curvas Características del Item (CCIs) son funciones[8] estrictamente crecientes $g_i(\xi)$ de los parámetros del rasgo latente $\xi \in R$.

Axioma (ii). Las CCIs son funciones de probabilidad:
$\lim_{\xi\to-\infty} g_i(\xi) = 0$ y $\lim_{\xi\to+\infty} g_i(\xi) = 1$.

Axioma (iii). Las respuestas al conjunto de *k* ítems que forman el test cumplen el principio de independencia local:

$$P\left[\left(X_{V_1} = x_{V_1}\right)\Lambda\left(X_{V_2} = x_{V_2}\right)\Lambda\ldots\Lambda\left(X_{V_K} = x_{V_K}\right)\right] = \prod_i g_i(\xi_v)^{x_{vi}}\left[1 - g_i(\xi_v)\right]^{1-x_{vi}}$$

A estos tres axiomas fundamentales añadimos el supuesto de suficiencia de la puntuación directa:

Axioma (iv). Dado un test de longitud k, la puntuación directa $R_v = \sum_{i=1}^{k} X_{vi}$ es un estadístico suficiente para ξ_v

Fischer (1995b) demuestra a partir de estos cuatro supuestos se deduce el modelo de Rasch, lo que formaliza en el siguiente teorema:

Teorema 1. Los axiomas (i), (ii) y (iii) junto a la suficiencia (iv) del estadístico R_v, implican un modelo de respuesta al ítem equivalente a un RM con CCIs:

$$f_i(\theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \tag{1}$$

[8] Notación: el parámetro personal o aptitud será denotado por ξ, el parámetro dificultad del ítem por δ_i y las CCIs por $g_i(\xi)$. Para obtener una métrica a nivel de intervalo será necesario realizar transformaciones monótonas de los parámetros (ϕ y ψ), definiéndose θ como $\phi(\xi)$ y β_i como $\psi(\delta_i)$, con CCIs $g_i(\xi) = f(\phi(\xi) - \psi(\delta_i)) = f(\theta - \beta_i)$.

Es decir, existe una función estrictamente creciente $\phi(\xi)= \theta$ y constantes β_i tales que todas las CCIs $f_i(\theta)$ se corresponden con la expresión (1).

Prueba. La suficiencia de la puntuación directa R (eliminamos el índice v para una exposición más clara) implica que, para cualquier vector de respuestas $\boldsymbol{x}$ y su correspondiente puntuación directa r, la función de probabilidad condicional $L(\boldsymbol{x}|\, r)$ satisface que

$$L(\boldsymbol{x}|\, r) = L(\boldsymbol{x}|\, \xi) \;/\; L(r|\, \xi) = c(\boldsymbol{x}) \qquad (2)$$

donde $c(\boldsymbol{x})$ es una constante que depende de $\boldsymbol{x}$ (y de r) y del conjunto de ítems, pero es independiente de ξ .Obsérvese que no se hacen supuestos acerca de la forma de la CCI, en concreto no se hacen supuestos acerca de si los ítems están o no caracterizados por un sólo parámetro.

Sea $\sum_{i=1}^{k} X_i$ = R suficiente para ξ, esto es que se cumpla la expresión (2) para un determinado vector de respuestas con $\sum_i x_i = r$ para un número *k de ítems* $i \geq 2$; definamos además $\boldsymbol{x}$ como un vector de respuestas con *a* (coeficiente de la transformación lineal) que satisfaga $x_1=1$ y $x_i=0$, $\boldsymbol{x}^{(1,i)}$ es para k>2 el vector de respuesta parcial $(x_2,\ldots, x_{i-1}, x_{i+1},\ldots, x_k)$, $\boldsymbol{y}$ como un vector de respuestas con *a* que satisface $y_1=0$ y $y_i=1$ e $y_l=x_l$, e $\boldsymbol{y}^{(1,i)}$ es para k>2 el vector de respuesta parcial $(y_2,\ldots,y_{i-1},y_{i+1},\ldots,y_k)$.

De acuerdo con los supuestos, $L(\boldsymbol{x}|\, \xi)= L(\boldsymbol{x}^{(1,i)}|\, \xi)g_1(\xi)[1-g_i(\xi)]$, con $L(\boldsymbol{x}^{(1,i)}|\, \xi):=1$ si k=2, de manera que (2) se convierte en

$$L(\boldsymbol{x}|r) = \frac{L(\boldsymbol{x}^{(1,i)}|\xi)g_1(\xi)[1-g_i(\xi)]}{L(r|\xi)} = c(\boldsymbol{x}) \qquad (3)$$

De forma similar, para el vector de respuestas $\boldsymbol{y}$,

$$L(\boldsymbol{y}|r) = \frac{L(\boldsymbol{y}^{(1,i)}|\xi)[1-g_1(\xi)]g_i(\xi)}{L(r|\xi)} = c(\boldsymbol{y}) \qquad (4)$$

y el cociente entre las expresiones (3) y (4):

$$\frac{L(\boldsymbol{x}^{(1,i)}|\xi)g_1(\xi)[1-g_i(\xi)]}{L(\boldsymbol{y}^{(1,i)}|\xi)[1-g_1(\xi)]g_i(\xi)}=\frac{c(\boldsymbol{x})}{c(\boldsymbol{y})} \tag{5}$$

que también es independiente de ξ. Obsérvese que por definición $L(\boldsymbol{x}^{(1,i)}|\ \xi) = L(\boldsymbol{y}^{(1,i)}|\ \xi)$ por lo que la expresión (5) implica que

$$\frac{g_1(\xi)[1-g_i(\xi)]}{[1-g_1(\xi)]g_i(\xi)}=\frac{c(\boldsymbol{x})}{c(\boldsymbol{y})}=d_i(\boldsymbol{x}) \tag{6}$$

que es independiente de ξ.

Hasta el momento no se ha realizado suposición alguna acerca de la escala de medida del rasgo latente ξ, por lo que, podemos elegir una transformación monótona conveniente de la escala $\xi \rightarrow \theta$, que en nuestro caso será la función logística:

$$f_1(\theta)=\frac{\exp(\theta)}{1+\exp(\theta)} \tag{7}$$

Reemplazando $g_1(\xi)$ y $g_i(\xi)$ en (6) por $f_1(\theta)$ y $f_i(\theta)$, respectivamente, y sustituyendo $f_1(\theta)$ por la expresión dada en (7) se obtiene la expresión del RM[9]

$$f_i=\frac{\exp(\theta\text{-}\beta_i)}{1\text{- }\exp(\theta\text{-}\beta_i)},\quad \text{con } \beta_i = \ln d_i(\boldsymbol{x}) \tag{8}$$

Por lo tanto, se demuestra que cualquier modelo que satisfaga los supuestos (i), (ii), (iii) y (iv) se puede transformar en un RM con parámetros para los ítems $\beta_1=0$ (parámetro de normalización) y β_i para i=2,...,k. En otras palabras, cualquier modelo que satisfaga estos cuatro supuestos será empíricamente un RM.

[9] La monotonía estricta de $g_1(\xi)$ supone la existencia de una función estrictamente monótona $\phi(\xi)$ para la definición de θ, tal que, $g_1(\xi)$ tome la forma (7).

Además, Fischer (1995b) demuestra que si R=$\sum_{i=1}^{k} X_i$ es un estadístico suficiente para el test formado por los ítems $I_{1,\dots,}I_k$, entonces R'=$\sum_{i=1}^{k-1} X_i$ será suficiente, a su vez, para el subtest formado por los ítems $I_{1,\dots,}I_{k-1}$. Es decir, de la suficiencia de R para el test completo, se deduce la suficiencia para ξ para todo subconjunto, no vacío, de ítems del test.

En definitiva, Fischer demuestra que de la suficiencia se deriva lógicamente el RM, modelo en el que existe una concatenación aditiva de los parámetros, por lo que, la transformación $\xi \rightarrow \theta$ mediante Rasch, implica que los parámetros θ y β_i entran en f_i en forma aditiva o sustractiva.

Cuando, además, se introduce la exigencia de concatenación aditiva de los parámetros θ y β_i de las CCIs, se deduce la medida en escala de intervalo, como demostró Pfanzagl (1971) a partir de la formulación del siguiente axioma (v) de concatenación aditiva y el lema siguiente:

Axioma (v). Existe un universo de ítems cuyas CCIs siguen la forma $f(\theta - \beta_i)$ para todo $\theta, \beta \in \mathbb{R}$, donde β es la dificultad del ítem, y *f(x)* es una función biyectiva continua creciente (por lo tanto, estrictamente creciente) $\mathbb{R} \rightarrow (0,1)$.

Lema 1. Si las probabilidades de respuesta correcta $P(+|\theta,\beta)= f(\theta - \beta_i)$ están fijadas para todo $\theta, \beta \in \mathbb{R}$, entonces del axioma (v) se deduce[10] que los parámetros θ y β son invariantes para las transformaciones lineales $a\theta + b_1$ y $a\beta + b_2$, siendo $a > 0$, b_1 y b_2 constantes arbitrarias, es decir, θ y β *son medidas en escalas de intervalo con unidad de medida común.*

Prueba. Supongamos un modelo de la TRI con CCIs de la forma $f(\theta - \beta_i)$. Esto implica que, para cualquier individuo con parámetro $\theta \in \mathbb{R}$, y para un ítem de dificultad $\beta \in \mathbb{R}$, la probabilidad $P(+|\theta,\beta)$

[10] Los axiomas (i) y (ii) están contenidos en (v).

está dada. Supongamos que existe otra representación compatible con estas probabilidades $P(+|\theta,\beta)$, con nuevos parámetros $k(\theta)$ y $l(\theta)$, y con CCIs $m[k(\theta)-l(\theta)]$, tal que

$$m[k(\theta)-l(\beta)]=f(\theta - \beta_i) \text{ para todo } \theta, \beta \in \mathbb{R}, \qquad (9)$$

donde k y l son una trasformación monótona estrictamente creciente de la escala en $\mathbb{R}$, y m está definido por (9). De lo que se sigue que:

$$k(\theta)-l(\beta)= m^{-1} \circ f(\theta - \beta_i), \qquad (10)$$

siendo "o" la concatenación de funciones. Para $\beta = 0$, la expresión (10) implica que $k(\theta)= m^{-1} \circ f(\theta) + l(0)$,

y si sustituimos $m^{-1} \circ f$ por h y l(0) por l_o, entonces

$$k(\theta)= h(\theta) + l_0, \qquad (11)$$

De forma similar, para θ=0 la expresión (10) conduce a

$$l(\beta)= -h(-\beta_i) + k_0 \qquad (12)$$

sustituyendo(12) y (11) en (10) se tiene que:

$$h(\theta) + l_0 + h(-\beta_i) - k_0= h(\theta-\beta) \qquad (13)$$

Para $\theta = \beta = 0$ se tiene que $h(0) = k_0-l_0= h_0$. Y de aquí,

$$h(\theta) - h_0 + h(-\beta_i) - h_0= h(\theta-\beta) - h_0 \qquad (14)$$

Para simplificar denotamos $h(x) - h_0$ por $h^*(s + t)$ con lo que queda la expresión:

$$h^*(s) + h^*(t) = h^*(s+t) \qquad (15)$$

con θ = s y $-\beta$= t. Esta es la ecuación funcional de Cauchy (Aczél, 1966), que bajo el supuesto de monotonía tiene la solución general $h^*(x)= ax$; a>0. Puesto que, $h^*(x) = h(x)- h_0$, se obtiene $h(x)= ax + h_0$. Que sustituyendo en (11) es $k(\theta)= a\theta + h_0 + l_0$, y en (12) es $l(\beta)= a\beta- h_0 + k_0$. De aquí que sean transformaciones admisibles de θ y β $a\theta + b_1$ y $a\beta + b_2$, con constantes arbitrarias a>0, b_1 y b_2.

Los cinco axiomas expuestos hasta el momento y el lema 1 llevan a la formulación de lo que se conoce como familia de modelos tipo Rasch

$$P(+|S_v, I_i) = \frac{\exp\left[a\left(\theta_v - \beta_i\right) + b\right]}{1 + \exp\left[a\left(\theta_v - \beta_i\right) + b\right]} \tag{16}$$

con constantes arbitrarias *a* y $b = b_1 - b_2$, donde *a* es la discriminación, que es común a todos los ítems, y *b* es una constante de escalamiento. Puesto que, cualquier cambio en la escala θ puede ser compensado con otro en la escala β, de los supuestos planteados se concluye que ambas escalas sean invariantes (o únicas) ante las transformaciones lineales positivas mencionadas anteriormente, lo que implica que los parámetros aptitud y dificultad son medidas en escala de intervalo.

A la demostración anterior puede objetarse que, como el supuesto (v), parece poco realista: la existencia de un universo ilimitado de ítems que se ajustan a unas determinadas CCIs para todo $\theta, \beta \in \mathbb{R}$. No obstante, Fischer (1995b) replantea el problema para el caso de que sólo existiera un universo finito de ítems, $I_1,...,I_k$ con parámetros β_i fijados, mostrando que incluso en este caso, la medida de los parámetros mantiene las propiedades de las escalas de intervalo para lo que formula el axioma y el lema siguientes:

Axioma (vii): Supuesto un universo finito de k ítems, con $k \geq 3$, cuyas CCIs toman la forma $f(\theta - \beta_i)$ para todo $\theta, \beta \in \mathbb{R}$; i=1,2,...,k, siendo *f* una función biyectiva continua creciente (esto es, estrictamente creciente) $\mathbb{R} \rightarrow (0,1)$. Entonces se puede demostrar que:

Lema 2: Todos los posibles cocientes $(\beta_j-\beta_i)/(\beta_q-\beta_p)$ son números racionales, donde θ es único en todos los puntos $\theta_{in} = \beta_i \pm n\Delta$, para n=0,1,2,...; i=1,...,k, siendo Δ un numero real tal que $\beta_j-\beta_i= w_{ji}\Delta$ para todos los pares (I_i,I_j) con pesos w_{ji}. Para todo θ_{in}, el parámetro θ es invariante para las transformaciones continuas estrictamente monótonas. Si existen al menos tres ítems, I_i, I_j, I_l, tales que el cociente $(\beta_j-\beta_i)/(\beta_l-\beta_j)$ sea un número irracional, la escala θ es una escala de intervalo.

Además de los anteriores supuestos, se han propuesto otros axiomas que permiten mostrar como de la suficiencia se derivan los RM, como la existencia de un estadístico minimal suficiente (Andersen, 1973), como se expresa en los dos siguientes axiomas:

Axioma (vii). Las CCIs con la forma $g(\xi)= g_i(\xi, \delta_i)$, donde δ_i es el parámetro escalar del ítem, y g una función continua $\mathbb{R}^2 \rightarrow (0, 1)$, tal que, para cualquier valor fijado $\delta_i \in \mathbb{R}$, $g(\cdot, \delta_i)$ es una función creciente biyectiva $\mathbb{R} \rightarrow (0, 1)$ y para cualquier valor fijado $\xi \in \mathbb{R}$, $g(\xi, \cdot)$ es una función decreciente biyectiva $\mathbb{R} \rightarrow (0, 1)$.

Axioma (viii). Dado que el test está formado por ítems $I_1,...,I_k$, debe de existir un estadístico T minimal suficiente no trivial para ξ que no dependa de los parámetros de los ítems $\delta_i,...,\delta_k$.

Si estos dos supuestos se combinan con el de independencia local (iii) puede demostrarse (ver Andersen, 1973) la existencia de transformaciones de la escala que son funciones continuas monótonas y estrictamente crecientes $\phi(\xi \rightarrow \theta)$ y $\psi(\delta \rightarrow \beta)$, ambas $\mathbb{R} \rightarrow \mathbb{R}$, tales que todas las CCIs tienen la forma RM dada en la expresión (1).

Andersen (1973) demuestra que se puede construir un estadístico minimal suficiente T independiente de δ_i, cuando $\delta_1 = \delta_2 = ... = \delta_k = \delta$. En este caso, la realización x_i con $i = 1,..., k$, sería la misma para todas la distribución de variables aleatorias. Si a este supuesto se aplica el resultado de Bahadur (1954) acerca de la unicidad de los estadísticos minimal suficientes, se puede seguir un razonamiento análogo al del teorema 1: T es alguna función biyectiva u de la puntuación directa R, $T=u(R)$, que implica su equivalencia con R. A partir de aquí se puede mostrar la suficiencia siguiendo los pasos dados en el primer teorema[11].

[11] Pfanzagl (1994) demostró de forma análoga que de la suficiencia se derivan los RM partiendo de la suposición de que las CCIs tienen la forma aditiva (v) para todo $\theta, \beta \in \mathbb{R}$, y de la existencia de un estadístico suficiente T no trivial para θ que es independiente de los parámetros de los ítems.

2.2 Transitividad conjunta

Guttman (1950) encuentra un gran problema de ambigüedad en el uso de las puntuaciones directas en los tests, puesto que comúnmente se observa que, distintos patrones de respuesta permiten obtener una misma puntuación, lo que lleva a situaciones como las siguientes: sujetos que en tests de aptitudes aciertan ítems de dificultad extrema, sin embargo, no aciertan otros de menor dificultad; o sujetos que en cuestionarios de personalidad o de actitudes responden afirmativamente a cuestiones que indican una postura extrema y, sin embargo, son más neutros en afirmaciones menos extremas, etc. Estas ambigüedades, se pueden eliminar si se exige a la escala lo que Guttman denominó transitividad conjunta, o la exigencia de un orden estricto tanto para las personas como para la dificultad de los ítems. Pues bien, se puede demostrar que el RM cumple con este principio, que se puede considerar como una aplicación de algunas de las exigencias para la medida de Campbell (1920) y que caracteriza a las escalas determinísticas de Guttman que formulamos a continuación:

Sean vDi y wDi las observaciones de que los sujetos S_v y S_w han resuelto correctamente el item I_i. Para los ítems I_i, I_j se tendrá una escala de Guttman cuando se cumpla que[12]

Si (vDi $\wedge \neg$ wDi $\wedge$ wDj) entonces vDj

Es decir, si el sujeto S_v resuelve correctamente el ítem I_i y el sujeto S_w no lo resuelve, pero si resuelve I_j, entonces el sujeto *v* también deberá resolver I_j. Por lo tanto, se establece un ordenamiento para sujetos e ítems.

Sea "iRj" donde "i es de mayor dificultad que j" en la escala latente, se sigue: $\neg$ vDi $\wedge$ vDj => iRj

Por lo que, aún desconociendo la localización precisa del sujeto S_v en la escala de aptitud, se sabe que estará delimitada por los

[12] $\wedge$ = "operador lógico y", $\neg$ ="operador no"

ítems I_i y I_j. Las escalas de Guttman exigen un ordenamiento de las dificultades de los ítems, independientemente de las aptitudes de los sujetos, en una escala común de dificultad y aptitud.

Aplicando un razonamiento análogo al marco probabilístico de la TRI, Roskam y Jansen (1984) plantean el siguiente axioma del que se derivan los RM:

Axioma (ix). P(iRj) es un ordenamiento estocásticamente consistente definido como la probabilidad $P(\neg \, vDi \wedge vDj \mid R_v = 1) = c$, donde R_v es la puntuación directa obtenida en los ítems I_i y I_j, y siendo c una constante independiente de v.

Fischer (1995b) observa que hay una equivalencia formal entre la propiedad de transitividad y la de suficiencia de la puntuación directa R_v, mostrando que de esta propiedad formalizada en el supuesto (ix) se deriva el RM. Para ello solo hay que seguir el procedimiento que se aplicaron en el apartado referente al principio de suficiencia.

2.3 Objetividad específica

Rasch (1967) formuló, como un principio científico general, el postulado de la objetividad específica (OE), que establece que *en un determinado y bien definido marco de referencia, las declaraciones científicas deben de ser generalizables*. Por declaraciones científicas se entienden las comparaciones entre objetos, y cualquier comparación debe de ser generalizable, más allá de la situación experimental concreta o instrumento (agente) con el que se ha realizado la comparación. Los resultados científicos que satisfagan este requisito de generalizabilidad serán denominados *comparaciones específicamente objetivas*. Rasch prefiere el término "objetivas" a "generalizables", aunque son conceptos intercambiables. Considera al OE un principio epistemológico general, aunque la OE en cada caso se refiera a un particular o específico marco de referencia, siendo necesario especificar el *tipo de comparaciones* y el *significado concreto de objetividad.*

Rasch (1960, 1967) utilizó para ilustrar cómo el significado de la objetividad de las comparaciones está basado en la definición conjunta, el ejemplo de las medidas de masa y fuerza en los mecanismos clásicos:

Sean los objetos de interés cuerpos rígidos, O_v, v= 1, 2,..., n con masas M_v. Si además existen condiciones experimentales (o agentes) denominados fuerzas F_i, que se van aplicando a cada una de las masas, de modo que se observan distintas aceleraciones A_{vi} . De acuerdo con el segundo axioma Newtoniano (fuerza = masa x aceleración), la aceleración observada será proporcional a la fuerza ejercida sobre el objeto e inversamente proporcional a su masa $A_{vi}= M_v^{-1} F_i$.

La comparación de dos objetos cualesquiera O_v y O_w con respecto a sus masas M_v y M_w se puede realizar mediante sus cocientes

$$\frac{A_{vi}}{A_{wi}} = \frac{M_v^{-1} F_i}{M_w^{-1} F_i} = \frac{M_w}{M_v} \qquad (17)$$

que son independientes de F_i. Aplicando una transformación logarítmica se obtiene una relación de concatenación aditiva: $\ln(A_{vi}/A_{wi})= \ln M_w - \ln M_v$. Esto significa que dos masas cualesquiera se pueden comparar sin necesidad de conocer las fuerzas F_i aplicadas. En otras palabras, la comparación entre dos masas cualesquiera es *generalizable* a todas las condiciones experimentales, con independencia de las aceleraciones a que hubieran dado lugar la aplicación de distintas fuerzas.

Puesto que la aceleración se puede establecer independientemente de la masa y de la fuerza (a partir de la distancia recorrida y el tiempo invertido), se hace posible el establecimiento de una definición de masa que permite su comparación (o medida) mediante la expresión (17).

Rasch (1967) aplicará este mismo esquema de pensamiento al problema de la comparación de aptitudes de cualesquiera dos sujetos, S_v y S_w , en relación con su respuesta a un ítem I_i de un test:

Para una muestra de sujetos, $S_1,\ldots,S_v,\ldots,S_n$, que responda a un conjunto de ítems del test $I_1,\ldots,I_i,\ldots,I_k$, se obtendrá una respuesta (una variable aleatoria R_{vi}) para cada par (S_v, I_i). Se supone que las respuestas son "observables", tanto directamente (caso determinístico), o indirectamente mediante el estudio de sus realizaciones globales (caso probabilístico). Además, la aptitud de cada sujeto S_v está totalmente caracterizada por sus reacciones (respuestas) a los ítems I_1, ..., I_k, por un parámetro latente ξ_v . Similarmente, la dificultad de cada ítem está totalmente caracterizado por un parámetro δ_i, y la variable reacción R_{vi} por el parámetro ρ_{vi} (todos los parámetros pueden ser vectores, pero para una mayor claridad expositiva restringiremos la demostración a parámetros escalares). Se considera que la comparación entre dos sujetos S_v y S_w con respecto a sus parámetros ξ_v y ξ_w será *siempre posible* una vez que se conozcan sus parámetros de reacción ρ_{vi} , ρ_{wi} y el resultado de la comparación será *siempre único*. El supuesto de la unicidad del resultado de la comparación implica que ρ_{vi} debe de ser alguna función F ("función de reacción") de ξ_v y $\delta_{i.}$

$$\rho_{vi} = F(\xi_v \text{ y } \delta_i) \qquad (18)$$

En Psicometría, la comparación de dos sujetos cualesquiera, S_v y S_w generalmente sólo se puede hacer mediante sus realizaciones R_{vi} y R_{wi} (variables de reacción), en lugar de con los parámetros que obviamente son inobservables ρ_{vi} y ρ_{wi}. La comparación entre sujetos, si se sigue el ejemplo de masa y fuerza, requerirá de la existencia de una función $W(R_{vi}, R_{wi})$ que lo haga posible. Para encontrar la función adecuada, habría que empezar buscando una función $U(\rho_{vi}, \rho_{wi})$ que nos devolviera el resultado en términos de ξ_v y ξ_w, puesto que, estos parámetros tienen toda la información pertinente contenida en las variables de reacción. Por lo tanto, asumamos la existencia de una función

$$U[F(\xi_v , \delta_{i.}), F(\xi_w , \delta_{i.})] = V(\xi_v, \xi_w) \qquad (19)$$

que, de acuerdo al postulado de OE, es independiente de δ_i y que denotaremos como $V(\xi_v, \xi_w)$. Para obtener el resultado buscado (18), es necesario establecer dos nuevos supuestos acerca de los parámetros y las funciones U y V:

Axioma (x). Si S_v[13] está caracterizado por el parámetro $\xi_w \in \mathbb{R}$ y el ítem I_i por el parámetro $\delta_i \in \mathbb{R}$, y existe una variable aleatoria R_{vi} para cada par (S_v, I_i) que está caracterizado por el parámetro de reacción $\rho_{vi} \in \mathbb{R}$; los parámetros de reacción están determinados por los parámetros respectivos de persona e ítem, $\rho=F(\xi, \delta)$, donde F es una función continua $\mathbb{R}^2 \rightarrow \mathbb{R}$, tal que $z = F(\cdot\ , y)$, para cualquier y fijado, es una función biyectiva creciente $\mathbb{R} \rightarrow \mathbb{R}$ y $z =(x\ , \cdot)$ para cualquier x fijado, una función biyectiva decreciente $\mathbb{R} \rightarrow \mathbb{R}$.

Axioma (xi). Existirá un comparador (función que permite la comparación) estrictamente monótono y continuo $U(\rho_{vi}\ , \rho_{wi})$, $\mathbb{R}^2 \rightarrow \mathbb{R}$, creciente en ρ_{vi} y decreciente en ρ_{wi}, para todo $\rho_{vi}, \rho_{wi} \in \mathbb{R}$, siempre que se satisfaga la ecuación funcional

$$U[F(x\ , z), F(y\ , z)] = V(x, y) \tag{20}$$

para todo x, y $z \in \mathbb{R}$, donde V es independiente de z.

Rasch (1977) dedujo la OE a partir de los supuestos (x) y (xi). No obstante, la OE puede derivarse de forma menos restrictiva (sin necesidad de asunciones tan exigentes), a partir de la teoría de ecuaciones funcionales, planteando el siguiente Lema:

Lema 3. los supuestos (x) y (xi) implican la existencia de funciones crecientes biyectivas continuas k, m, u, y f, todas $\mathbb{R} \rightarrow \mathbb{R}$, tales que

$$U(\rho_{vi}, \rho_{wi}) = V(\xi_v, \xi_w) = u[k(\xi_v)\text{-}k(\xi_w) = u(\theta_u\text{-}\theta_w)$$

[13] Supuestos dos sujetos S_v y S_w con parámetros $\xi_v \neq \xi_w$ si tuvieran la misma variable reacción con respecto al mismo ítem I_i implicaría que $\rho_{vi} = \rho_{wi}$, lo que entraría en contradicción con la unicidad de la comparación entre sujetos, puesto que, si dos parámetros personales son iguales deben serlo también los parámetros de reacción. Esta es la motivación de la suposición de monotonía respecto a la función z.

$$F(\xi_v , \delta_{i.}) = f[k(\xi_v)- m(\delta_i)]= f(\theta_u-\beta_i)$$

donde $\theta_u = k(\xi_v)$ y es $\beta_i = m(\delta_i)$.

Prueba. La monotonía estricta de F en ambas variables implica que, fijada z, F(y,z)=w pueda invertirse la relación, es decir, debe de existir una función continua G tal que G(w,z) = y, donde G es una función monótona estrictamente creciente en ambas variables. Si incluimos este principio en la ecuación funcional de (xi) tenemos:

$$U[F(x,z),w] = V[x, G(w,z)] \tag{21}$$

que es una *ecuación funcional asociativa generalizada* (Aczél, 1966). Bajo el supuesto de continuidad y monotonía realizada anteriormente, la solución continua general de esta ecuación toma la forma

$$\begin{aligned} &U(x,y) = l[a(x) - b(y)], \quad F(x,y) = a^{-1}[k(x) - m(y)] \\ &V(x,y) = l[k(x) - h(y)], \quad G(x,y) = -h^{-1}[-m(x) - b(y)] \end{aligned} \tag{22}$$

donde a, b, h, k, l, m son funciones crecientes biyectivas $\mathbb{R} \rightarrow \mathbb{R}$

Insertando (22) en la ecuación funcional (ix) se tiene que[14]

$$l[a \circ a^{-1}[k(x) - m(z)]-b \circ a^{-1}[k(y) - m(z)]] = l[k(x)-h(y)]$$
$$b \circ a^{-1}[k(y) - m(z)] = h(y) - m(z)$$

y sustituyendo s= k(y), y = $k^{-1}(s)$, t= m(z), queda

$$b \circ a^{-1}(s-t) = h \circ k^{-1}(s) - t \tag{23}$$

para t=0 en (23), se tiene que es

$$b \circ a^{-1}(s) = h \circ k^{-1}(s) \tag{24}$$

de forma similar para s=0 en (23), se tiene que

$$b \circ a^{-1}(-t) = h \circ k^{-1}(0) - t$$

[14] Utilizamos el símbolo "o" para indicar la concatenación de funciones

que es equivalente a

$$b \circ a^{-1}(t) = t + c \qquad (25)$$

donde la constante $c = h \circ k^{-1}(0)$. Además, con $p= a^{-1}(t)$, se tiene que $t=a(p)$, de lo que se sigue de la expresión (25) que

$$b(p)= a(p) + c \qquad (26)$$

Similarmente, (24) y (25) implican que

$$h(q) = k(q) + c \qquad (27)$$

Insertando (26) y (27) en (22), con $u(x)= l(x\text{-}c)$, se obtiene que

$$U(x,y) = l[a(x) - a(y)-c] = u[a(x) - a(y)]$$
$$F(x,y) = a^{-1}[k(x)-m(y)]$$
$$V(x,y) = l[k(x) - h(y)-c] = u[k(x) - k(y)]$$

Volviendo a sustituir este resultado en la ecuación funcional del supuesto (xi), se tiene que es:

$$U(F(x,z), F(y,z)) = u[a(F(x,z))- a(F(y,z))] =$$
$$u[a \circ a^{-1}(k(x) - m(z)) - a \circ a^{-1} (k(y) - m(z))] =$$
$$u[k(x) - k(y)] = V(x,y)$$

Denotando $a^{-1}(x)$ por $f(x)$, la función de respuesta F toma la forma final

$$F(x,y) = f[k(x) - m(y)]$$

siendo el comparador $V(x,y) = u[k(x) - k(y)]$

Retornando a la notación original donde $k(\xi) =\theta$ y $m(\delta) = \beta$, se observa que la función de reacción se puede escribir como

$F(\xi_v,\delta_i) = f(\theta_v - \beta_i)$, y el comparador

$$V(\xi_1, \xi_2) = u (\theta_v - \theta_w)$$

que es el resultado que se buscaba.

Rasch(1977) demuestra que la OE, junto con la continuidad y la monotonía, implican una determinada "aditividad en el rasgo latente" (o "sustractividad" con la notación aquí utilizada) en la función de respuesta F y en el comparador U. La aditividad en los parámetros F y U se obtiene aplicando una transformación de las escalas k(ξ) y m(δ) a las escalas originales latentes, con lo que se establecen nuevas escalas para θ y β. Aunque las escalas originales ξ y δ sean sólo ordinales, por la propia naturaleza de las nociones psicológicas de "aptitud" y "dificultad", las transformaciones k(ξ) y m(δ) son específicas e invariantes (únicas) para las transformaciones lineales dadas en el lema 1, lo que supone que las escalas aditivas resultantes θ y β tienen las propiedades de las escalas de intervalo.

Obsérvese que la derivación de la aditividad del rasgo se ha realizado para cada ítem del test, en lugar de para el total del test. Rasch consideraba que la comparación debe de ser independiente de la elección del conjunto de ítems (agentes), lo que haría posible que con un simple agente fuera suficiente para hacer la comparación.

Debe observarse que la derivación tanto de la aditividad de los parámetros θ y β en la función de reacción *f*, como la de θ_v y θ_w en la función comparador U, es general, sin haberse especificado ningún modelo concreto de la TRI. Este principio general se aplica tanto al caso de fuerza y masa en física, como a la aptitud y dificultad en Psicología.

Una cuestión diferente es la de concretar cuales son los modelos de la TRI compatibles con el principio OE. Comprobaremos a continuación cómo, los modelos dicotómicos de la familia RM, cumplen con esta exigencia para la medida científica. Para obtener este resultado, se requiere hacer una pequeña modificación en el supuesto de OE, debiéndose introducir un nuevo postulado que expresaremos como: dentro de la TRI, la mayoría de las conclusiones acerca de los ítems o de los sujetos están basadas en la probabilidad o verisimilitud de las respuestas observadas (con formato de respuesta dicotómica en esta demostración), por lo que es necesario que el comparador U sea una función de probabilidad (condicional o incondicional) de las respuestas observadas. Formalicemos este nuevo supuesto (xii):

Sea A={(+ +), (+ -), (- +), (- -)} el espacio muestral de todos los posibles patrones de respuesta para dos sujetos S_v y S_w a un item I_i, y sean B y C subconjuntos de este espacio muestral $C \subset B \subseteq A$. Entonces:

Axioma (xii). Para dos sujetos cualesquiera S_v y S_w que responden a un ítem I_i, debe de existir un comparador $U:(0, 1) \times (0, 1) \rightarrow \mathbb{R}$ de la forma $U[g(\xi_v,\delta_i), g(\xi_w,\delta_i)]$, donde $g_i(\xi) = g(\xi,\delta)$ sean las CCIs, y donde U satisfaga las siguientes condiciones:

(a) U(x, y) es una función continua, monótona estrictamente creciente en *x* para todo *y*, siendo estrictamente monótona decreciente en *y* para todo *x*.

(b) $U [g(\xi_v,\delta_i), g(\xi_w,\delta_i)] = V(\xi_v, \xi_w)$ es independiente de δ_i, para todo $\xi_v, \xi_w, \delta_i \in \mathbb{R}$ (es decir, la comparación de S_v con S_w basada en el ítem I_i es OE).

(c) Principio de verosimilitud: U debe de ser o una función de probabilidad no condicionada P(B), o bien una función de probabilidad condicionada $P(C \mid B)$.

Teorema 2. Los supuestos (vii), (iii) y (xii) implican la existencia de funciones crecientes continuas biyectivas k y l, que permiten las transformaciones $\xi \rightarrow \theta$ y $\delta \rightarrow \beta$, respectivamente , ambas $\mathbb{R} \rightarrow \mathbb{R}$, tales que las CCIs para todo S_v e I_i tomen la forma de la funciones de RM (16). Una vez realizada la transformación, θ y β serán medidas invariantes en escala de intervalo.

Prueba: Considerando las respuestas de dos sujetos S_v y S_w a un ítem I_i, nos preguntamos si existe o no alguna función de probabilidad no trivial P(B) con $B \subseteq A$, o alguna función de probabilidad condicional $P(C| B)$ con $C \subset B \subseteq A$, que satisfaga los supuestos del teorema 2. Rasch (1968) responde a esta pregunta mostrando que existen dos y sólo dos funciones de probabilidad condicionada $P[(+ -) \mid (+ -) \vee (- +)]$ y $P[(- +)| (+ -) \vee (- +)]$ que además son complementarias.

Evaluemos las distintas posibilidades:

En primer lugar contemplemos la posibilidad de que la función U sea una función de probabilidad no condicionada P(B). Para una menor complejidad de notación denotemos $g(\xi_v, \delta_i)$ por p_{vi}.

Si B consistiera en un conjunto con un elemento A, por ejemplo, $B=\{+ -\}$, entonces $P(B) = p_{vi}(1-p_{wi})$ debería tener la forma $V(\xi_v, \xi_w)$, donde V es alguna función de los parámetros personales e independiente de δ_i. Sin embargo, en el caso de que ambos parámetros personales fueran iguales $\xi_v = \xi_w$, se tendría que $p_{vi}(1-p_{wi}) = p_{vi} - (p_{vi})^2$, resultado que entra en contradicción con el supuesto de independencia de δ_i. De forma similar, puede mostrarse que esto ocurre con todos los casos de un elemento, y también con los casos de tres elementos, al ser éstos complementarios de los elementos simples, por lo que únicamente quedarían como posibles candidatos los casos con pares de elementos.

Considérese el conjunto con un par B= {(- +),(+ +)}, en este caso, cuando son iguales los parámetros personales, se tiene que $P(B)= (1-p_{vi})p_{wi} + p_{vi}\ p_{wi} = p_{wi} = V(\xi_v,\ \xi_w)$, lo que entra en contradicción con los supuestos al ser p_{wi} independiente de ξ_v. En todos los demás casos con pares, se encuentran contradicciones similares, con lo que queda rechazada la posibilidad de encontrar una función U de probabilidad no condicionada que cumpla con los supuestos planteados en el teorema 2.

Analicemos ahora si U puede ser una función de probabilidad condicionada P(C| B). En primer lugar, es obvio que B no puede ser un elemento simple, puesto que C no puede ser un conjunto vacío, con lo que B tendría que ser un par o un trío. Existen 24 posibles tríadas B con subconjuntos $C \subset B$, siendo muchos idénticos, por ser permutaciones de los 3 elementos considerados. Comencemos analizando el caso en que B = {(- -),(- +), (+ +}, con C={(- +)}

$$P(C|B) = \frac{(1-p_{vi})p_{wi}}{(1-p_{vi})(1-p_{wi})+(1-p_{vi})p_{wi}+p_{vi}p_{wi}} \tag{28}$$

Si $\xi_v = \xi_w$ tenemos que:

$$P(C|B) = \frac{p_{vi}(1-p_{vi})}{1-p_{vi}(1-p_{vi})} \tag{29}$$

resultado que debería de ser independiente de δ_i. Sin embargo, se observa que $p_{vi}(1-p_{vi}) = p_{vi} - (p_{vi})^2$, que entra en contradicción con la suposición de independencia, y esto mismo ocurre con el resto de tríadas, quedando por lo tanto todas ellas eliminadas.

En cuanto a los conjuntos de pares, se pueden formar seis casos para B, en cuatro de ellos el descarte es inmediato: {(- -),(- +)}, {(- -),(+ -)}, {(- +),(+ +)}, {(+ -), (+ +}, en todos ellos B estaría completamente determinado por la respuesta de uno de los sujetos, por lo que cada una de las P(C|B) equivale a una de las probabilidades incondicionales p_{vi}, $1\text{-}p_{vi}$, p_{wi}, $1\text{-}p_{wi}$, ninguna de las cuales puede ser independiente de δ_i. Al realizar, por último, la evaluación de los pares {(- -), (+ +)} y {(+ -), (- +)}, se observa que para el primero de estos dos pares cuando C={(+ +)}, se tiene

$$P(C|B)=\frac{p_{vi}p_{wi}}{(1\text{-}p_{vi})(1\text{-}p_{wi})+p_{vi}p_{wi}}=\frac{\dfrac{p_{vi}}{1\text{-}p_{vi}}\ \dfrac{p_{wi}}{1\text{-}p_{wi}}}{1\text{-}\dfrac{p_{vi}}{1\text{-}p_{vi}}\ \dfrac{p_{wi}}{1\text{-}p_{wi}}} \qquad (30)$$

que para $\xi_v = \xi_w$, sería monótona estrictamente creciente en p_{vi}, por lo que no sería independiente de δ_i. Además, en el caso de C={(- -)} encontraríamos una contradicción análoga.

Por lo tanto, sólo queda la opción de B= {(+ -), (- +)}. Cuando C={(+ -)}, la probabilidad condicionada es:

$$P(C|\ B)=\frac{p_{vi}(1\text{-}p_{wi})}{p_{vi}(1\text{-}p_{wi})+(1\text{-}p_{vi})p_{wi}} \qquad (31)$$

que bajo el supuesto de independencia de δ_i., implica que

$$\frac{p_{vi}(1\text{-}p_{wi})}{(1\text{-}p_{vi})p_{wi}}=\frac{V(\xi_v,\xi_w)}{1\text{-}V(\xi_v,\xi_w)} \qquad (32)$$

La expresión (32) caracteriza la comparación de dos sujetos S_v y S_w basada en un ítem I_i y es formalmente análoga a la expresión (6), en la que se comparaban dos ítems a partir de la respuesta de un sujeto, por lo que se obtiene el mismo resultado que con el teorema 1. Cuando la función $k_l(\xi)$ es estrictamente monótona para todos los ítems I_i, se sigue que:

$$\ln\left(\frac{g_1(\xi)}{1-g_1(\xi)}\right)=k_1(\xi) \tag{33}$$

Haciendo uso de (33) para el ítem I_i la expresión (32) se puede escribir como

$$k_i(\xi_v)-k_i(\xi_w)=\ln\left(\frac{V(\xi_v,\xi_w)}{1-V(\xi_v,\xi_w)}\right) \tag{34}$$

que debería de ser independiente de I_i. De aquí que, para cualquier elección para un "ítem estándar", I_0, se tiene que

$$k_i(\xi_v) - k_i(\xi_w) = k_o(\xi_v) - k_0(\xi_w),$$
$$k_i(\xi_v) - k_o(\xi_v) = k_i(\xi_w) - k_0(\xi_w) \tag{35}$$

para todo ξ_v, $\xi_w \in \mathbb{R}$. Puesto que, el primer miembro de la ecuación (35) depende de ξ_v, mientras que el segundo depende de ξ_w, esta ecuación es de hecho una constante que depende de los ítems I_i e I_0. A la parte derecha de la ecuación la denotamos por $-\beta_{0i}$, con lo que (35) se puede expresar como:

$$k_i(\xi_v) = k_o(\xi_v) - \beta_{0i} \tag{36}$$

Sustituyendo (36) en (33)

$$\ln\left(\frac{g_1(\xi)}{1-g_1(\xi)}\right)= k_0(\xi_v)-\beta_{0i}$$

$$g_i(\xi_v)=\frac{\exp(k_0(\xi_v)-\beta_{0i})}{1+\exp(k_0(\xi_v)-\beta_{0i})}$$

que es la expresión (1) del RM con $\theta_v= k_0(\xi_v)$ y $\beta_i = \beta_{0i}$, como caso particular de la *familia de modelos de Rasch* (16).

Queda así demostrada la OE en esta familia de modelos, junto a la propiedad de unicidad de la representación ante las transformaciones lineales $a\theta + b_1$ y $a\beta + b_2$ con $a>0$, ya demostrada anteriormente en la prueba del Lema 1.

3. La familia de modelos de Rasch y extensiones

Las propiedades del RM y de la familia de modelos de Rasch se pueden derivar también en sus extensiones politómicas y en sus desarrollos unidimensionales y multidimensionales. Una excelente revisión de estos modelos se puede ver en la obra de Fisher y Molenaar (1995).

Una de las extensiones más fructíferas y prometedoras del RM es la que se conoce como el modelo logístico lineal. Aunque en sus primeras formulaciones se puede considerar un modelo unidimensional, permite hacer una descomposición lineal de los parámetros dificultad de los ítems en determinados parámetros "básicos", en función del tipo concreto de aplicación: operaciones cognitivas necesarias para la resolución de los ítems, situaciones de evaluación de los sujetos, o para la evaluación de los tratamientos. Las aplicaciones más comunes han consistido en intentar evaluar la dificultad de las operaciones cognitivas como elementos del proceso de resolución, pero también se puede utilizar para medir los efectos de las condiciones experimentales sobre la dificultad de los ítems, o el impacto de los distintos tratamientos educativos, o de las terapias clínicas, etc. Además del modelo logístico lineal y de sus extensiones, también nos ocuparemos de otras extensiones del RM en el que éste se combina con el análisis de clase latente y que ha mostrado su utilidad en la modelización de las actitudes y de los constructos de personalidad. Por último, analizaremos los conocidos como modelos componenciales, modelos que guardan grandes similitudes con los dos anteriores y que han mostrado una gran utilidad para la modelización de los constructos psicológicos.

3.1 El modelo logístico lineal

Para introducir al modelo logístico lineal o LLTM (*Linear Logistic Test Model*) recordemos que el RM modela la probabilidad de que una persona S_j resuelva un ítem I_i en términos de diferencia en el parámetro de aptitud θ_j y del parámetro dificultad del ítem β_i. La probabilidad de respuesta correcta esta condicionada a los parámetros aptitud y dificultad y se puede formalizar como:

$$P\left(+\middle|\theta_j,\beta_i\right)=P_{ij}=\frac{\exp(\theta_j-\beta_i)}{1+\exp(\theta_j-\beta_i)} \tag{37}$$

El RM define el parámetro reacción ρ_{ij} como la diferencia entre los parámetros personal (aptitud) y de ítem (dificultad), relación que se obtiene al transformar el RM en un modelo *logit*

$$\rho_{ij} = \ln\left(\frac{P_{ij}}{1-P_{ij}}\right)=\theta_j-\beta_i \tag{38}$$

A partir de una matriz completa de respuestas "ítems x personas" se puede concebir un modelo logit completo bifactorial: un factor es el ítem con parámetro β_i, y el otro el factor persona con parámetro θ_j. El modelo bifactorial se extiende a un modelo multifactorial al añadir factores adicionales L, M, ..., con sus respectivos parámetros λ_l, μ_m, ..., que caractericen los efectos de las condiciones experimentales que influencian las probabilidades de respuesta:

$$\rho_{ijlm}= \theta_j-\beta_i+\lambda_l+\mu_m+.... \tag{39}$$

que es el modelo multifactorial completo sugerido inicialmente por Rasch (1965) y que engloba al modelo LLTM que posteriormente elaboró Fischer (1973).

En el LLTM el parámetro RM dificultad del ítems β_i se descompone en múltiples parámetros:

$$\beta_i = \sum_{l=1}^{p} w_{il}\alpha_l + c \qquad (40)$$

donde α_l con l = 1,...., p, son los "parámetros básicos" del ítem, w_{il} son los pesos de cada parámetro básico, y c es una constante (de normalización). Los parámetros básicos, según la aplicación concreta, pueden ser las modalidades de las condiciones experimentales u otros factores que influencien la dificultad del ítem. Estos factores se supone que son fuentes de diferencias en dificultad entre ítems.

De la ecuación (40) se sigue que, para cualquier par de ítems (I_i, I_k), la diferencia β_i - β_k se puede explicar por los parámetros básicos α_l con pesos w_{il}, de modo que

$$\beta_i - \beta_k = \sum_{l=1}^{p} (w_{il} - w_{kl})\alpha_l \qquad (41)$$

De aquí que, para que el modelo ajuste, sea necesario que todos los factores que contribuyan a las diferencias entre ítems deban de tenerse en cuenta. Mientras que aquellos factores que influyan por igual en todos los ítems, se contemplan en la constante de normalización c.

Fischer (1995c) presenta como una prueba que avala el uso del LLTM, el que éste pueda ser asimilado a un modelo de contrastada calidad métrica como es el Bradley-Terry-Luce (BTL). El modelo BLT (e.g. Luce, 1959) tiene una larga tradición en Psicología matemática y permite explicar el resultado de una comparación entre dos objetos O_i y O_j de un conjunto $\{O_1,...,O_n\}$, por medio de las probabilidades $P(O_i \succ O_j)$, donde $O_i \succ O_j$ significa que en una comparación concreta se prefiere O_i a O_j, y en el que todas las comparaciones se asume que son estocásticamente independientes. Pues bien, Fischer muestra como el RM, y consiguientemente el LLTM, se pueden considerar como ejemplos de comparaciones BLT incompletas entre personas e ítems.

El modelo LLTM habitualmente es considerado un RM con restricciones impuestas a los parámetros de los ítems. No obstante, Fischer (1997) observa que el LLTM es más general en su estructura formal, por lo que el RM sería un caso especial de LLTM donde la matriz de pesos, **W**, es diagonal. Además, el LLTM es más general que el RM, puesto que la función de verosimilitud, las ecuaciones de estimación y la unicidad de condiciones para el RM, hacen que este sea un caso especial de LLTM.

En cuanto a la aplicabilidad del LLTM, éste es más restrictivo que el RM, ya que exige: (a) que el RM ajuste a los datos, y (b) que los parámetros de los ítems satisfagan la ecuación lineal (40). De aquí que no sea fácil encontrar aplicaciones del LLTM, al exigir un gran esfuerzo en la planificación de la investigación y en la cuidadosa elaboración de los ítems del test. En cuanto a la estimación de los parámetros del modelo, afortunadamente hoy día contamos con buenos programas de ordenador como el LPCM-Dos o LPCM-WIN 1.0 (desarrollados Seliger y Fisher, 1994; Fischer y Ponocny-Seliger, 1998).

A pesar de los problemas para obtener una adecuada bondad de ajuste se han intentado y realizado múltiples aplicaciones prácticas. Las aplicaciones más tempranas del LLTM se centraron en explicar la dificultad de los ítems, en términos de dificultad de las operaciones cognitivas involucradas en el proceso de resolución de esos ítems, al modo sugerido por Scheiblechner (1972). Este autor, utilizando un test de razonamiento lógico, fue pionero en aplicar el análisis de regresión para estimar los parámetros de los ítems, obteniendo un análisis de la complejidad cognitiva de las proposiciones lógicas presentadas a los sujetos en forma gráfica, a partir de tres fuentes de dificultad (negación, disyunción y asimetría). Así, por ejemplo, si un ítem requería la operación lógica de la "disyunción", el peso correspondiente w_{il} era la unidad, y cero en caso de no requerir la operación de disyunción. Spada (1976) estudió diversas tareas a partir de sus mecanismos elementales, y Kubinger (1979) estudió las facetas que se requiere para superar un curso de Estadística. Otras investigaciones se han centrado en tests de razonamiento (Embretson, 1985; Hornke y Habon, 1986; Smith, Kramer, y

Kubiak, 1992; Real, Olea, Ponsoda, Revuelta y Abad, 1999; López y Elosua, 2002, Gorin, 2003).

Fischer (1997) recoge el proceso de elaboración de matrices no-verbales similares a las matrices progresivas de Raven seguido por Formann en 1973 en el que se consideraban tres factores o facetas como determinantes de las operaciones cognitivas. Aunque el ajuste sólo fue exitoso parcialmente, la aplicación del LLTM sirvió como herramienta para mejorar la construcción del test. Un ejemplo más reciente, en el que se muestra la utilidad del LLTM en la construcción de tests, la proporciona Gittler (1991) , quien desarrolló un nuevo test utilizando cubos, diseñado para medir la aptitud espacial, y que ha mostrado una excelente bondad de ajuste.

La aplicación para tests de aptitudes, está limitada a dominios de ítems donde el investigador dispone de buenas hipótesis acerca del conjunto de operaciones cognitivas relevantes que se requieren para resolver cada ítem. La no especificación de la matriz **W** del sistema lineal, obviamente nos conduce a errores sistemáticos en la estimación de los parámetros, que puede afectar seriamente al ajuste del modelo. De hecho, los problemas que los investigadores han tenido al intentar aplicar los modelos LLTM a tests tradicionales se explica porque éstos no se han desarrollado partiendo de un modelo que especifique las operaciones que permitan tener un universo de ítems bien definidos, no obstante, parece ser una camino prometedor para la construcción de tests (Fischer, 1995c). Además, el LLTM también puede ser utilizado para la detección de los efectos de la práctica entre sesiones sucesivas de evaluación. Los métodos tradicionales, tanto en la teoría clásica como en la TRI, confunden los efectos de la práctica con la dificultad de los ítems. El LLTM, sin embargo, es capaz de describir y predecir la dificultad de los ítems en función de una teoría formalizada de la complejidad cognitiva de los ítems: los modelos dinámicos para el aprendizaje.

Entre las aplicaciones más usuales del LLTM, encontramos que se ha aplicado con éxito a investigaciones inter-culturales, específicamente para la comparación de la disponibilidad de distintas funciones cognitivas en diversas culturas (Piswanger, 1975; Fischer y Forman, 1982; Mohammadzadeh-Koucheri, 1993).

El LLTM requiere no sólo homogeneidad en los ítems en el sentido del RM, sino también homogeneidad respecto a las estrategias cognitivas empleadas por las personas. Por ello, la heterogeneidad en las estrategias utilizadas, es una de las principales causas de la falta de ajuste del modelo. Para intentar resolver este problema Mislevy (1988) propone incorporar una variación aleatoria a los parámetros de los ítems β_i , con la misma estructura genérica $\mathbf{w}_i$, asumiendo una determinada función de densidad normal multivariante. Van Maanen, Been y Sijtsma (1989) proponen como solución hacer el ajuste para los distintos subgrupos de sujetos. Esta idea da lugar a los modelos de distribución de mixturas como el propuesto por Rost (1990), que se aborda en el siguiente epígrafe.

Por último, la extensión del LLTM conocida como modelo logístico lineal para el cambio, es una buena alternativa para su aplicación en estudios que utilizan diseños longitudinales y de medidas repetidas. Según explican Fisher y Selinger (1997) el LLTM puede fácilmente reinterpretarse como un modelo para el estudio del cambio y para la medida de los efectos del tratamiento, que pasa a denominarse *Linear Logistic Model with Relaxed Assumptions* (LLRA). En este desarrollo del LLTM se relaja el supuesto de unidimensionalidad, para hacer posible el estudio de los cambios producidos al pasar un mismo ítem a los mismos sujetos en distintas ocasiones (dos o más). El LLRA ha sido ideado pensando en su aplicación en investigaciones psicológicas en los campos educativos o clínicos, donde es frecuente observar una heterogeneidad en los ítems (o en los síntomas). Por ejemplo, considérese el caso del seguimiento de la maduración cognitiva en los niños: un conjunto de ítems apropiados para la evaluación intelectual durante el desarrollo, deberá contener ítems que valoren diferentes factores de inteligencia, lo que habitualmente lleva a la construcción de distintas escalas unidimensionales, que podría hacerse mediante un modelo LLTM. Sin embargo, una solución más realista sería la aplicación del LLRA, que permitiría evaluar tanto la maduración cognitiva como las diferencias existentes, tanto intra como intergrupo (por ejemplo, entre niños y niñas).

3.2 Modelos de distribución de mixturas

El análisis de clase latente o LCA (*Latent Class Analysis*) es una técnica multivariante que permite identificar clases distintas de individuos en una escala psicométrica (Lazarfield y Henry, 1968). Los individuos de una determinada clase se supone que tendrán un comportamiento similar que debe diferir del de otras clases de individuos. El LCA tiene un gran interés, puesto que permite modelar constructos *fuertes,* como la inteligencia o las aptitudes, sino también otros más *débiles*, permitiendo, por ejemplo, identificar subtipos de personalidad (Gangestad y Snyder, 1985).

En los modelos TRI, generalmente se parte del supuesto de que los datos provienen de una única población y que todos los miembros responden a los ítems utilizando los mismos procesos y estrategias, que se puede caracterizar mediante la función de respuesta al ítem. Sin embargo, el descubrimiento de clases al aplicar LCA, implicaría que los sujetos responden de forma diferente a los mismos ítems, con lo que ya no sería adecuado situar a todos los sujetos en un único continuo psicométrico.

La suposición básica de los modelos de distribución de mixturas, o distribuciones mezcladas, es que la distribución de la variable aleatoria observada debe ser descrito por más de una función de probabilidad. Las probabilidades estarán condicionadas para la variable *mezcla,* que es una variable constituida por una mixtura de distribuciones.

El LCA puede entenderse como una distribución simple de mixturas discretas para la respuesta a los ítems, que se basa en dos suposiciones, la independencia local y la constancia de las probabilidades de respuesta en todos los individuos que forman la clase (Rost y Davier, 1995):

$$P(X=x)=\sum_{c=1}^{c}\pi_c\prod_{i=1}^{k}\pi_{i|c}^{x_i}(1-\pi_{i|c})^{1-x_i} \tag{42}$$

aquí hay dos tipos de parámetros, los *parámetros de tamaño de la clase* π_c=P(c) o proporciones de la mezcla de la mixtura total, y las probabilidades condicionadas del ítem $\pi_{i|c}=P(x_i = 1 \mid c)$.

En el modelo de Rost (1990), conocido por el acrónimo MIRA (*MIxed RAsch model*), la función de probabilidad de la variable X (distribución de mixturas) tiene la forma

$$P(X=x)=\sum_{c=1}^{c}\pi_c P(X=x \mid c) \tag{43}$$

donde las probabilidades condicionadas $P(X = \boldsymbol{x} \mid c)$ de la distribución de mixturas están definidas por el modelo de Rasch, pudiéndose formalizar como

$$P(\theta)=\sum_{c}\pi_c \frac{\exp(\theta_{jc}-\beta_{ic})}{1+\exp(\theta_{jc}-\beta_{ic})} \tag{44}$$

donde θ_{jc} es el nivel en el rasgo de la persona *j* de la clase *c*, β_{ic} es la dificultad del ítem *i* en el grupo *c, y* π_c es el parámetro tamaño de la clase o proporción de la mixtura o mezcla.

Por lo tanto, el MIRA hacer una integración entre el LCA y el RM, al basarse en el supuesto de que cada componente de la distribución total, o mezcla de distribuciones, se distribuye como un RM, con lo que se hace posible la identificación de las distintas subpoblaciones (clases) dentro de la población general, cada una caracterizada por sus propios patrones de respuesta. También existe una extensión de este modelo para datos dicotómicos a datos politómicos, que el lector puede consultar en Davier y Rost (1995a).

Para la aplicación práctica del LCA y del MIRA, se han desarrollado programas de ordenador específicos para los modelos de Rost como el MIRA (Rost y Davier, 1992) o el WINMIRA (Davier, 1994; Davier y Rost, 1995b) que se pueden utilizar para modelar tanto datos dicotómicos como politómicos. Estos modelos, en su versión dicotómica, se han aplicado a tareas de razonamiento espacial (Mislevi et al., 1990; 1991; Embretson, 2004), y en su versión politómica se han realizado aplicaciones prácticas en cuestionarios de actitudes (Rost y Georg, 1991) y de personalidad (Rost, Carstersen y Davier, 1997).

3. 3 Modelos componenciales

Entre los modelos multidimensionales de la TRI elaborados a fin de poder modelar la complejidad de los procesos cognitivos, al permitir descomponer los parámetros dificultad y/o los parámetros aptitud, destacan los conocidos como modelos componenciales, que también están considerados como una extensión de la familia RM.

El Modelo Multicomponente de Rasgo Latente (MLTM), propuesto inicialmente por Withely (1980), fue desarrollado para la medida de diferencias individuales en los componentes del procesamiento subyacente, en tareas de aptitudes complejas. El MLTM es un modelo multidimensional conjuntivo en que la realización de la tarea depende de los componentes de dificultad de las tareas y de los componentes de aptitud de las personas. Una generalización de este modelo, conocida como el modelo general de componentes de rasgo latente (GLTM; Embretson, 1984) permite conectar las características estimulares de la tarea con los componentes de dificultad de esa tarea. Además, es común que en los estudios cognitivos se manipulen las características de los estímulos para controlar la dificultad de los procesos específicos, por lo que el modelo generalizado permite una mejor conexión con la teoría cognitiva que el modelo original MLTM.

Los modelos MLTM y GLTM se han utilizado para la búsqueda de fuentes de diferencias individuales, pero quizás, las aplicaciones más extensas se han dado en el campo de investigación de la validación de constructo.

El modelo MLTM especifica una relación conjuntiva entre la respuesta a un ítem y la respuesta a un componente de procesamiento subyacente. Es decir, los componentes subyacentes deben ser puestos en práctica para la correcta resolución de la tarea. De manera que, la probabilidad de resolver correctamente la tarea *U* es una probabilidad condicionada de la aptitud y la dificultad que vendría dada por el producto de las probabilidades de resolver las tareas individuales:

$$P\left(U_{ijT}=1\middle|\theta_j,b_i\right)=(s\text{-}g)\prod_k P(U_{ijk}=1|\theta_{jk},b_{ik})+g \tag{45}$$

donde θ_{jk} es la aptitud de la persona *j* en el componente *k*, b_{ik} es la dificultad del ítem i en el componente k, U_{ijk} es la respuesta de la persona *j* en el componente k-éximo del ítem *i*, U_{ijT} es la respuesta de la persona *j* en la tarea total T para el ítem *i*, *g* es la probabilidad de resolver el ítem por azar y *s* es la probabilidad de aplicar el resultado o solución del componente para resolver la tarea.

A su vez, la utilización de los componentes de procesamiento depende de la aptitud de las personas y de la dificultad de la tarea. Los parámetros personales y los de los ítems, en el MLTM, están incluidos en un modelo logístico que contiene las componentes de aptitud θ_{jk} y de dificultad b_{ik}, de modo que el modelo MLTM completo se puede expresar como:

$$P\left(U_{ijT}=1\middle|\theta_j,b_i\right)=(s\text{-}g)\prod_k \frac{\exp(\theta_{jk}\text{-}b_{ik})}{1+\exp(\theta_{jk}\text{-}b_{ik})}+g \tag{46}$$

Una inspección de esta expresión revela que las probabilidades de los componentes están especificadas por un RM.

Por ejemplo, el establecimiento de analogías verbales, se asume que involucra a dos componentes generales, las reglas de construcción de esas analogías y la evaluación de las posibles alternativas de respuesta. Así, en el ítem que a continuación ponemos como ejemplo, se trata de evaluar la relación entre dos términos (ej: evento y memoria), para después descubrir el término que guarda una relación análoga con el ítem de prueba (ej: Fuego):

Evento - Memoria
Fuego: _____ (1. Manguera 2. Cenizas 3. Agua 4. Calor)

Para realizar correctamente la anterior tarea se exige (a) descubrir la analogía o regla subyacente entre los antecedentes y (b) reconocer la alternativa correcta, para lo cual hay que plantearse ¿qué queda cuando el fuego se extingue?.

El MLTM proporciona la probabilidad de estimar la aptitud para resolver correctamente la tarea, descomponiendo ésta en sus dos componentes de dificultad: probabilidad de construir la regla correctamente y probabilidad de evaluar correctamente las posibles alternativas de respuesta. Existe una variante de este modelo que permite modelar estrategias secuencialmente dependientes: en esta extensión del modelo sólo se activa una estrategia cuando la inmediata anterior, una vez ordenadas en un orden jerárquico de dificultad, haya resultado exitosa. En este caso, la probabilidad de resolver los ítems se representa mediante la suma de probabilidades condicionadas a las estrategias.

El GLTM, como ya se ha indicado, es una generalización del MLTM que permite establecer restricciones en las dificultades de los ítems, de modo similar a la generación del LLTM a partir del RM. Pero, a diferencia del LLTM en el que sólo se modeliza la respuesta total al ítem, el GLTM permite establecer restricciones en los componentes, lo que permite estudiar la respuesta a los componentes de dificultad que se asume están relacionados con distintos procesos cognitivos. El modelo GLTM se puede formalizar como:

$$P\left(U_{ijT}=1\middle|\theta_j,\alpha_k,\beta_k\right)=(s-g)\prod_k \frac{\exp(\theta_{jk}-\sum_m \beta_{mk} v_{imk}+\alpha_g)}{1+\exp(\theta_{jk}-\sum_m \beta_{mk} v_{imk}+\alpha_g)}+g \qquad (47)$$

donde v_{imk} representa la dificultad del ítem *i* en el estímulo factor *m* del componente *k*, β_{mk} es el peso del factor complejidad de *m* en el componente *k*, α_k es la constante de normalización para el componente *k*, *g* es la probabilidad de resolver el ítem por azar y *s* es la probabilidad de aplicar los resultados de los componente para resolver la tarea.

McCollam, Schmidt y Embretson (1999) recientemente han presentado una pequeña modificación del GLTM denominada *General Component Latent Trait Model for Covert Processes* (GLTM-CP). Esta variante es más flexible en cuanto a su aplicabilidad, pudiéndose aplicar cuando se carece de los datos correspondientes a cada subtarea. El GLTM-CP permite estimar diferencias individuales en procesos totalmente encubiertos a partir de un ítem estándar, mediante el uso de técnicas para datos incompletos (*missing data*). En este caso, la dificultad del proceso subyacente se modela en dos etapas, partiendo de las características de los ítems: en primer lugar, se modelan las características de la complejidad del estímulo que definen cada ítem; posteriormente, en el ámbito las componentes, la respuesta al ítem se modela a partir de las aptitudes y de las dificultades pronosticadas para los procesos.

Para los modelos componenciales, la estimación de los parámetros de los componentes se debe a Maris (1995), quien extendió el modelo de Withely (1980) utilizando el algoritmo EM para datos incompletos (Dempster, Laird y Rubin, 1977). Para el modelo GLTM-CP, la identificación del modelo se obtiene mediante la imposición de restricciones a los componentes de dificultad de los ítems (Embretson, 2004).

Los modelos componenciales se han aplicado con éxito a estudios de validación de constructo, como analogías verbales y de clasificaciones (Embretson, 1985; Embretson et al., 1986; Embretson, 1997), problemas matemáticos (Bertrand et al., 1993), matrices progresivas (Embretson, 1994), ítems de vocabulario (Janssen et al., 1991) y tareas espaciales (Embretson, 2004).

Para Embretson (1997) el uso de estos modelos componenciales aportan importantes ventajas a los estudios de validación de constructo, puesto que proporcionan un conjunto completo de parámetros para representar, tanto a los distintos aspecto del constructo (parámetros personales), como las diferentes tareas (parámetros de los ítems).

La elaboración de nuevos tests para la investigación de aptitudes, a partir de los modelos de procesamiento cognitivo, permite analizar por separado los dos aspectos que constituyen la validez de constructo: la representación del constructo y el espacio nomotético. La representación del constructo se centra en la identificación de los procesos subyacentes, en el conocimiento de estructuras y las estrategias que están involucradas en la realización de las tareas y cuyo conocimiento debe de estar sustentado en la modelización matemática. Para evaluar el espacio nomotético, que nos informa de la utilidad del test y nos permite conocer su capacidad predictiva, será necesario conocer las relaciones con otras medidas del constructo, para lo que será preciso recurrir a técnicas como las matrices multimétodo-multirrasgo, las ecuaciones estructurales, etc.

En el proceso de construcción de tests, el uso de modelos componenciales puede resultar muy útil para el diseño de ítems para la medida de componentes cognitivos específicos. Por ejemplo, existen al menos dos modos de aislar o medir un único componente, o bien eliminando los ítems con bajos b_{ik} en el resto de componentes durante la selección de los ítems, o bien, y probablemente esta sea una solución, utilizar los factores de complejidad estimular como guía para enunciar los ítems.

Otras aplicaciones de los modelos componenciales es la búsqueda de diferencias entre grupos como, por ejemplo: la búsqueda de diferencias en la resolución de problemas matemáticos, en función del género, como hacen Bertrand et al.(1993), en una investigación en la que se examinó el impacto de diferentes estilos de vida entre los estudiantes (horas viendo televisión, dedicadas a la lectura, etc.) sobre los componentes y sobre la tarea global.

La revisión de los distintos modelos que se presentan en este capítulo, nos hace estar esperanzados en cuanto al desarrollo presente y futuro de nuevos procedimientos psicométricos de medida que sean más realistas, en cuanto a que se generen a partir de modelos cognitivos, más rigurosos, en cuanto a que se ajustaran a los requisitos de la medición científica y más válidos, al compaginar su utilidad con una adecuada representación del constructo. Afortunadamente, en los últimos años se viene observando un uso creciente en los estudios de validez mediante modelos de la familia RM, como los aquí tratados, y de otras aproximaciones multidimensionales (ver Van der Linden y Hambleton, 1997).

4. Epílogo

Generalmente los problemas de la medición y los modelos propuestos desde la TRI, como forma de superar estos problemas, se han tratado de forma separada de los de la validez de los instrumentos psicométricos. Una de las pocas excepciones han sido las contribuciones de Embretson en las que se muestran como la TRI y fundamentalmente la modelización multicomponencial, abren el camino a una nueva dimensión de validez: *la representación del constructo*, que tiene que ver con la identificación de los mecanismos que subyacen a la realización de la tarea. Esta perspectiva se une a la tradicional *nomotética* de la validación del constructo, que se refiere a la red de relaciones que se establece entre un test y otras medidas. Aquí hemos querido tratar de forma conjunta ambos problemas, el de la justificación y/o calidad de la medición y el de la validez de los instrumentos psicométricos, puesto que pensamos carece de sentido cuestionarse acerca de la validez de un determinado procedimiento psicométrico, si antes no se ha dado respuesta a cuestiones más básicas y fundamentales como, por ejemplo, si los ítems forman o no una escala, si esta escala es o no de naturaleza cuantitativa, si hay definida una métrica, qué operaciones son admisibles en la escala, si la escala es lineal, etc.

Confiamos en que este texto haya arrojado luz y pueda contribuir a entender la complejidad que entraña establecer medidas válidas y rigurosas de los constructos psicológicos. Estas cuestiones ya se plantearon en los trabajos que Campbell, Thurstone y Fisher entre otros, desarrollaron en los años 20 del siglo pasado, aunque prácticamente fueron ignorados durante décadas por los diseñadores y constructores de tests y de lo que Michell, en sus múltiples escritos, culpabiliza a Stevens. Pero no es responsabilidad de Stevens, puesto que mostró sus dudas sobre la calidad métrica de la mayoría de las escalas psicológicas, el que gran parte de los psicólogos *comenzaran una huída hacia adelante* y consideraran legitimado el uso cuantitativo de las puntuaciones brutas o directas de los tests, de cualquier test, bastando con que mostrasen ser suficientemente fiables (esto es, que tengan una aceptable homogeneidad interna) y válidos en el sentido que se consideren útiles para un determinado propósito de evaluación.

Afortunadamente, las contribuciones posteriores de Luce, Tukey, Rasch, Fischer, Wright, Embretson, entre muchos otros, incidiendo en la exigencia de una medición rigurosa y fundamentada, como condición y requisito para el establecimiento de su validez, finalmente han dado sus frutos como, por ejemplo, queda patente en un reciente monográfico sobre la validez publicado en la revista Metodología de las Ciencias del Comportamiento (Vol 5(2), 2004). En este monográfico, prácticamente la mitad de las contribuciones giran en torno a los problemas aquí abordados y que podemos resumir en que toda medida psicométrica debe de comenzar con una definición clara del objeto de la medida y de un posicionamiento claro ante la definición del constructo (Lohman, 1994). Posteriormente habrá que seleccionar un modelo que permita representar adecuadamente los datos (Bond, 2004; Embretson, 2004), modelo que además ha de tener la mayor calidad métrica posible, constituyendo una excelente opción los modelos de la familia de Rasch y sus extensiones que han sido expuestos en la presente monografía.

REFERENCIAS BIBLIOGRÁFICAS

Aczél, J. (1966). *Lectures on functional equations and their applications*. New York: Academic Press.

Allen, M.J. and Yen, W.M (1979) *Introduction to Measurement Theory*. Belmont, California: Wadsworth, Inc.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (AERA, APA, NCME). (1985). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (AERA, APA, NCME). (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Psychological Association.

American Psychological Association (APA) (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, *51*, 201-238.

American Psychological Association, American Educational Research Association and National Council on Measurement in Education (APA, AERA, NCME). (1966). *Standards for educational and psychological tests and manuals*. Washington, D.C.: Author.

American Psychological Association, American Educational Research Association and National Council on Measurement in Education

(APA, AERA, NCME). (1974). *Standards for educational and psychological tests*. Washington, D.C.: American Psychological Association.

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measure*ment, *10*, 67-78.

Anastasi, A. (1954). *Psychological testing*. New York: McMillan.

Anastasi, A. (1986). Evolving concepts of tests validation. *Annual Reviews of Psychology*, *37*, 1-15.

Anastasi, A. (1988). *Psychological Testing (6ª Ed.)* Macmillan Publishing Company: New York.

Andersen, E. B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26*, 31-44.

Andrich, D. (1995) Models for measurement. Precision and the non-dichotomization of grades responses. *Psychometrika*, *60*, 7-26

Andrich, D. (1996) Measurement criteria for choosing among models for graded responses. En A. von Eye and C. C. Clogg (Eds). *Analysis of categorical variables in developmental research* (pp. 3-35) Orlando: Academic Press

Angoff, W. H. (1988). Validity: An evolving concept. En H. Wainer and H.Braun (Eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Assessment Systems Corporation (1995). *The Rasch model item calibration program. User´s manual for the MicroCAT testing system*. St. Paul, Minnesota.

Bahadur, R. R. (1954). Sufficiency and statistical decisión functions. *Annals of Mathematical Statistics*, *25*, 423-462.

Barthélémy, J. P. (1984) *About the asymetries of similarity judgments: An ordinal point of view. Trends in Mathematical Psychology*. E. Degreef and V. Buggenhaut (Eds.). Amsterdam: North Holland. pp.177-191.

Bertrand, R., Dupuis, F.A., and Garneau, M. (1993). *Effets des caractéristiques des items sur le rôle des composantes impliquées dans la performance en résolution de problèmes mathématiques écrits: une étude de validité de construit* (Rapport FCAR). Québec (Canada): Université Laval.

Binet, A. and Simon, Th (1908). Le dévelopement de l'intelligence chez les enfants. *Année Psychologique*, *14*, 1-94.

Bond, T. G. (2004). Validation and assessment: a Rasch measurement perspective. *Metodología de las Ciencias del Comportamiento*, *5*, 181-196.

Bookstein, A. (1992). Informetric Distributions, Parts I and II. *Journal of the American Society for Information Science*, *41*, 368-88.

Bookstein, A. (1996). Informetric distributions III. Ambiguity and randomness. *Journal of the American Society for Information Science*, *48*, 2-10.

Borsboom, D. and Mellenbergh, G. J. (2004). Why psychometrics is not pathological: A comment on Michell . *Theory & Psychology*, *14*, 105-120

Bradley, R. A. (1954a) Rank analysis of incomplete block designs II. Additional tables for the method of paired comparisons. *Biometrika*, *41*, 502-537.

Bradley, R. A. (1954b) Incomplete block rank analysis: on the appropriateness of the model for a method of paired comparisons. *Biometrics*, *10*, 375- 390.

Bradley, R. A. (1955) Rank analysis of incomplete block designs III. Some large sample results on estimation and power for a method of paired comparisons. *Biometrika*, *42*, 450-470

Bradley, R. A. and Terry, M. E. (1952) The rank analysis of incomplete block designs. *Biometrika*, *39*, 324-345

Bridgman, P. W. (1922). Dimensional analysis. New Haven: Yale University Press

Brogden, H. E. (1977). The Rasch model, the law of comparative judgement and additive conjoint measurement. *Psychometrika*, *42*, 631-634

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, and discriminant validity. *American Psychologist*, *15*, 546-553.

Campbell, D. T. and Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psycological Bulletin*, *56*, 81-105.

Campbell, N. R. (1920). *Physics: The elements*. Cambridge: Cambridge University Press

Campbell, N. R. (1928). *An account of the principles of measurement and calculation*. London: Logmans Green

Campbell, N. R. (1957). *What is Science?* New York: Dover Publications (Reprint de Campbell, 1920)

Carnap, R. (1966). *Philosophical foundations of Physics*. New York: Basic Books

Cook, T. D. and Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Coombs, C. H. (1952). *A theory of psychological scaling*. Eng. Res. Bull. 34. Ann Arbor: University of Michigan Press

Coombs, C. H., Dawes, R. M. and Tversky, A. (1970) *Mathematical psychology: An elementary introduction*. Englewood Cliffs, NJ: Prentice-Hall, 1970.

Coombs, C. H., Raiffa, H. and Thrall, R. M. (1954). Some views on mathematical models and measurement theory. *Psychological Review 61*, 132-144

Cronbach, L. J. (1949/1984). *Essentials of psychological testing*. New York: Harper.

Cronbach, L. J. (1971). *Test validation*. En R.L. Thorndike (Ed.), Educational Measurement (2ª ed.). Washington, D.C.: American Council on Education.

Cronbach, L. J. (1975). Five decades of public controversy over mental testing. *American Psychologist*, *30*, 1-14.

Cronbach, L. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, theory, and pubic policy* (pp. 147-171). Urbana , IL : University of Illinois Press.

Cronbach, L. J. and Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana: University of Illinois Press.

Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: John Wiley.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psichological Bulletin*, *52*, 281-302.

Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement*, *10*, 94-96.

Daniel, M. H. (1999). *Behind the scenes: Using new measurement methods on DAS and KAIT*. En Embretson, S. E and Hershberger, S.L. (Eds) New Jersey: LEA, Publishers

Davier, v M. (1994). *WINMIRA. Program manual.* Kiel: IPN

Davier, v M. and Rost, J. (1995a). Polytomous Mixed Rasch Models. En G. H. Fischer e I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Davier, v M. and Rost, J. (1995b). *WINMIRA: Windows Mixed Rasch Model análisis* [Computer program]. Kiel, Netherlands: Institute for Science Education.

Debreu, G. (1960). Topological methos in cardinal utility theory. En K.J. Arrow, S. Karlin and P. Suppes (Eds.). *Mathematical methods in the social sciences*. Stanford: Standford University Press

Dempster, A. P, Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood estimating with incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *Series B*, *39*, 1-38.

Doignon, J. P. (1984). *Generalizations of interval orders. Trends in Mathematical Psychology*. 209- 217. E. Degreef and V. Buggenhaut (Eds.). Amsterdam: North Holland

Doignon, J. P., Ducamp, A. and Falmagne, J. C. (1984) On realizable biorders and the biorder dimension of a relation. *Journal of Mathematical Psychology 28*, 73-109

Droste, M. (1987a). Ordinal scales in the theory of measurement. *Journal of Mathematical Psychology 31*, 60-82

Droste, M. (1987b). Classification and transformation of ordinal scales in the theory of measurement. En *Progress in Mathematical Psychology*, I. E.E. Roskam and R. Suck. (Eds.). Amsterdam: Noth-Holland. pp. 47-966

Ellis, B. (1968). *Basic concepts of measurement.* Cambridge: Cambridge University Press

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span, *Psychological Bulletin*, *93*, 179-197.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika*, *49*, 175-186.

Embretson, S. E. (1985). *Test design: Developments in psychology and psychometrics*. New York: Academic Press.

Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. Reynolds (Ed.), *Advances in cognitive assessment: An interdisciplinary perspective* (pp. 107-135). New York: Plenum.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, *8*, 341-349

Embretson, S. E. (1997). Multicomponent Response Models. En W.J. van der Linden and R.K. Hambleton (Eds), *Handbook of modern item response theory*. New York: Springer-verlag

Embretson, S. E. (2004). Application of two IRT models for contruct validation to issues about spatial ability. *Metodología de las Ciencias del Comportamiento*, *5*, 159-180.

Embretson, S. E. and Hershberger, S. L. (Ed.) (1999). *The new rules of measurement.* Mahwah, New Jersey: LEA

Embretson, S. E., Schneider, L. M. and Roth, D. L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement*, *23*, 13-32.

Fechner, G. T. (1860). *Elemente der Psychophysik. Leipzig: Breitkopf und Härtel.* See, Fechner, G. T. (1966). *Elements of Psychophysics.* Volume I. New York: Holt, Rinehart and Winston.

Fischer, G. (1968). *Psychologische testtheorie.* Bern, Switzerland: Huber

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359-374.

Fischer, G. H. (1995a). Some neglected problems in IRT. *Psychometrika*, *4*, 459-487.

Fischer, G. H. (1995b). Derivations of the Rasch model. En G. H. Fischer and I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Fischer, G. H. (1995c). The Linear Logistic Test Model. En G. H. Fischer and I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Fischer, G. H. (1997). Unidimensional Linear Logistic Rasch Models. En W. J. van der Linden and R. K. Hambleton (Eds), *Handbook of modern item response theory*. New York: Springer-verlag

Fischer, G. H. and Forman, A. K. (1982). Some applications of logistic latent trait models with linear constraints on the parameters. *Applied Pychololological Measurement*, *4*, 397-416.

Fischer, G. H. and Molenaar, I. W. (1995) *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Fisher, G. H. and Ponocny-Seliger, E. (1998). *Structural Rasch Modeling. Handbook of the usage of LPCM-WIN 1.0*. Netherlands: ProGAMMA.

Fischer, G. H. and Selinger, E. (1997) Multidimensional Linear Logistic Models for Change. En W. J. van der Linden and R. K. Hambleton (Eds), *Handbook of modern item response theory*. New York: Springer-verlag

Fisher, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error and by the mean square error. *Monthly Notices of the Royal Astronomical Society*, *53*, 758-770.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London*, *222*, 309-368.

Gangestad, S. and Snyder, M. (1985). To carve nature at its joints: On the existence of discrete classes in personality. *Psychological Review*, *92*, 317-349.

Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist*, *27*, 197-222.

Gittler, G. (1991) *Dreidimensionaler Würfeltest (3DW). Ein Rasch-skalierter Test Zur Messung des räumlichen Vorstellungsvermögens*. Weinheim: Beltz Test.

Gorin, J. S. (2003). Cognitive and psychometric modeling of text-based reading-comprehension items form the gre-v items. Dissertation abstracts international: *Section B: The Sciences and Engineering*, *64*, 1544.

Guilford, J. P. (1946). New standards for tests evaluation. *Educational and Psychological Measurement*, *6*, 427-439.

Gulliksen, H. (1950a). *Theory of mental tests* (reeditado en 1987). New York: John Wiley.

Gulliksen, H. (1950b). Intrinsic validity. *American Psychologist*, *5*, 511-517.

Guttman, L. (1950). *The basis for scalogram analysis. In Stouffer et al. Measurement and Prediction, Volume 4*. Princeton N.J.: Princeton University Press, 60-90.

Helmholtz, H. von. (1887). Zählen und Messen, erkenntnistheoretisch betrachtet. En *Philosophische Aufsätze Eduard Zeller gewidmet*. Reprinted in Helmholtz, H.von. (1895) Gesammelte Abhandlungen, Vol 3

Helmholtz, H. von. (1977). Numering and measuring from an epistemological viewpoint. En R.S. Cohen and Y. Elkana (Eds.) Hernan von Helmholtz, *Epistemological Writtings*. Dordrecht: Reidel

Hempel, C. G. (1952). *Fundamentals of concept formation in empirical science*. Int.Enc. Unified Science, 2, 7. Chicago: University Chicago Press

Hölder, O. (1901). Die axiome der quantität und die lehre vom mass. Berichte der sächsischen gesselschaft der wissenschaften, *mathem. phys. klasse*, *53*, 1-64

Hornke, L. F. and Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement*, *10*, 369-380.

Hunter, J. E., Schmidt, F. C. and Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage

Jackson, D. N. (1970). A sequential system for personality scale development. In C. D. Spielberger (Ed.), *Current topics in clinical and community psychology*, (Vol. 2, pp. 61–81). New York: Academic Press.

Janssen, R., Hoskens, M. and DeBoeck, P. (1991). A test of Embretson's multicomponent model on vocabulary items. In R. Steyer and K. Widaman (Eds.), *Psychometric Methodology*. Stuttgart, Germany: Springer-Verlag.

Kolmogorov, A. N. (1950). *Foundations of the Theory of Probability*. New York: Chelsea Publishing.

Kubinger K. D. (1979). Das Problemlöseverhalten bei der statistischen Auswertung psychologischer Experimente. Ein Beitrag Hochschuldidaktischer Forschung. Zeitschrift für *Experimentelle und Angewndte Psychologie*, *26*, 467-495.

Koslow, A. (1982). *Quantity and Quality: Some aspects of Measurement*. En Preceeding of the Philosophy of Science Association, PSA 1, 183-198

Krantz, D. H., Luce, R. D., Suppes, P. and Tversky, A (1971). *Foundations of measurement, Vol 1*. New York: Academic Press

Kuhn, T. S. (1961). The function of measurement in modern Physics. *Isis 52*, 161-190

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University Chicago Press

Kyburg, H. E. (1984). *Theory and measurement. Cambridge Studies in Philosophy*. London: Cambridge University Press

Lazarsfeld, P. F., and Henry, N. W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.

Levy, P. (1937). *Theorie de l´addition des variables aleatoires*. Paris: Masson

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*,635-694 (Monograph Supplement 9).

Loève, M (1976). *Teoría de la Probabilidad.* Tecnos.

Lohman, D. F. (2004). Conceptions of validity as framework for the validation of ability constructs. *Metodología de las Ciencias del Comportamiento*, *5*, 143-158.

López Jauregui, A. y Elosua Oliden, P. (2002). Formulación y validación de un modelo logístico lineal para la tarea de adición y sustracción de fracciones y números mixtos. *Psicothema*, *14*, 802-809.

Lord, F. M. and Novick M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, Mass: Addison-Wesley.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, New Jersey: Lawrence Erlbaum

Luce, R. D. (1956). Semi-orders and theory of utility discrimination. *Econometrica*, *24*, 178-191

Luce, R. D. (1959). *Individual choice behavior. A theoretical analysis*. New York: Wiley

Luce, R. D. and Tukey, F. W. (1964). Simultaneous conjoint measurement: a new type of fundamental measurement. *Journal of Mathematical Psychology*, *1*, 1-27

Mach, E. (1960). *The science of mechanics*. La Salle: Open Court

Maris, E. M. (1995). Psychometric latent response models. *Psychometrika*, *60*, 523-547.

McCollam, K. M. Schmidt and Embretson, S. E. (1999). *Decomposing age-related differences in cognition using The General Latent Trait Model for Covert Processes. (GLTM-CP).* Paper presented at The Psychometric Society Meeting, Lawrence, KS.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012-1027.

Messick, S. (1989/1993). Validity. En R.L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: American Council on Educational/Macmillan.

Messick, S. (1995). Validity of Psychological Assessment. Validation of inferences from person's responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749.

Michell, J. (1999) *Measurement in Psychology: Critical History of Methodological Concept*. Cambridge University Press.

Michel, J. (2000). Normal science, pathological science and psychometrics . *Theory & Psychology*, *10*, 639-667.

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. Applied *Psychological Measurement*, *12*, 281-296.

Mislevy, R. J. and Verhelst, N. (1990). Modeling item responses when differnt subjects employ different solution strategies. *Psychometrika*, *55*, 195-215.

Mislevy, R. J. Wingersky, M. S., Irvine, S. H. and Dann, P. L. (1991). Resolving mixtures of strategies in spatial visualisation tasks. *British Journal of Mathematical and Statistical Psychology*, *44*, 265-288.

Mohammadzadeh-Koucheri, F. (1993). *Interkultureller Vergleich mit einer variierten Form des Matrizentests von Forman*. Vienna: University of Vienna.

Monjardet, B. (1978). Axiomatiques et propriétés des quasi-orders. *Mathematics, Sciences Humaines*, *63*, 51-82

Narens, L. (1980). On qualitative axiomatizations of probability. *Journal of Philosophical Logit*, *9*, 143-151

Narens, L. (1985). *Abstract measurement theory*. Cambridge, Mass: MIT Press

Pfanzagl, J. (1968). *Theory of measurement*. New York: Wiley

Piswanger, K. (1975). *Interkulturelle Vergleiche mit dem Matrizentest von Formann* (Tesis doctoral). Vienna: University of Vienna.

Pfanzagl, J. (1971). *Theory of measurement*. New York: J. Wiley.

Pfanzagl, J. (1994). On item parameter estimation in certain latent trait models. En G.H. Fischer and D. Laming (Eds.), *Contributions to Mathematical Psychology, Psychometrics, and Methodology*. New York: Springer-Verlag.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: MESA Press

Rasch, G. (1961). *On general laws and meaning of measurement in psychology*. Proceeding of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 4, 321-333. Berkeley: University of California Press

Rasch, G. (1963). *The Poisson process as a model for a diversity of behavioral phenomena*. International Congress of Psychology: Washington, DC

Rasch, G. (1965). *Statistiks seminar* (Apuntes tomados por J. Stene) Copenhagen: Department of Statistics, University of Copenhagen.

Rasch, G. (1967). An informal report an a theory of objectivity in comparison. En L.J.Th. Van der Kamp and C.A.J. Vlek (Eds.), *Psychological Measurement theory*. Leyden: University of Leyden.

Rasch, G. (1968). *A mathematical theory of objetivity and its consequences for model construction*. En Report from European Meeting on Statistics, Econometrics and Management Sciences. Amsterdam

Rasch, G. (1969). *Models for description of the time-space distribution of traffic accidents*. Symposium on the Use of Statistical Methods in the Analysis of Road Accidents. Organization for Economic Cooperation and Development Report 9

Rasch, G. (1977). On specific objetivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, *14*, 58-94

Real, E., Olea, J., Ponsoda, V., Revuelta, J. y Abad, F. J. (1999). Análisis de la dificultad de un test de matemáticas mediante un modelo componencial. *Psicológica*, *20*, 121-134.

Roberts, F. S. (1979), *Measurement Theory, with Applications to Decisionmaking, Utility, and the Social Sciences*, Addison-Wesley, Reading, MA.

Roskam, E. E. (1983). Allgemeine datentheorie. En H. Feger and H. Bredenkamp (Eds.) *Messen und Testen*. 1-135 Vol. 3 de la Serie Forschungsmethoden der Psychologie der Enzyklopädie der Psychologie. Göttingen: Hogrefe

Roskam, E. E. and Jansen, P. G. (1984). A new derivation of the Rasch model. *Trends in Mathematical Psychology*. 293-307. E. Degreef and V. Buggenhaut (Eds.). Amsterdam: North Holland

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item análisis. *Applied Psychological Measurement*, *3*, 271-282.

Rost, J., Carstersen C. and Davier v. M. (1997). Applying the Mixed Rasch Model to Personality Questionnaires. En J. Rost and R.Langeheine (Eds.), *Applications of Latent Trait and Latent Class Models in the Social Sciences*. Waxmann: Münster.

Rost, J. and Davier, v., M. (1992). *MIRA. A PC-programm for the Mixed Rasch model- User Manual*. Kiel: Institute for Science Education (IPN).

Rost, J. and Georg, W. (1991). Altenative Skalierungs-möglichkeiten zur klassischen Testtheorie am Beispiel der Skala "Jugendzentrismus". *ZA-Information*, *28*, 52-75.

Roy, B. (1980). *Preference, indiference, incomparabilié*. Documents du LAMSADE, 9. Université de Paris-Dauphine

Rulon, P. J. (1946). On the validity of educational tests. Harvard *Educational Review*, *16*, 290-296.

Santisteban, C. (1990). *Psicometría: Teoría y práctica en la construcción de tests*. Madrid: Ediciones Norma

Santisteban, C. (2003). *Los Métodos y el Progreso de la Ciencia*. Madrid: Ediciones UNED

Santisteban, C. y Alvarado J. M. (2001) *Modelos Psicométricos*. Madrid: Ediciones UNED

Scheiblechner, H. (1972). The learning and solving of complex reasoning items. *Zeitschrift für Experimentelle und Angewandte Psychologie*, *3*, 456-506.

Schwager, K. W. (1991). The representational theory of measurement: an assessment. *Psychological Bulletin*, *110*, 618-626.

Seliger, E. and Fisher, G. H. (1994). *Program description - LRSMG, LRSM, LLTM, and LPCM, with applictions to sacale analysis and measuring change*. Vienna: Department of Psychology of the University of Vienna.

Scott, D. (1964). Measurement models and linear inequalities. *Journal of Mathematical Psychology*, *1*, 233-247

Scott, D. and Suppes, P. (1958). Foundational aspects of theories of measurement. *Journal of Symbolic Logit*, *23*, 113-128

Smith, R. M., Kramer, G. A., and Kubiak, A. T. (1992). Components of difficulty in spatial ability test items. En M. Wilson (Ed.), *Objective measurement: Theory into practice, Vol. I.* Norwood, NJ: Ablex Publishing corporation.

Spada, H.(1976). *Modelle des Denkens und Lernes.* Berne: Huber.

Stevens, S. S. (1939). On the problem of scales for the measurement of psychological magnitudes. *Journal for the Unification of Science*, *9*, 94-99.

Stevens, S. S. (1946). On the theory of scales and measurement. *Science*, *103*, 667-680.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. En S. S. Stevens (Ed.) *Handbook of Experimental Psychology*. New York: Wiley

Suck, R. (1987). Approximation theorems for conjoint measurement models. En *Progress in Mathematical Psychology*, I. E.E. Roskam and R. Suck.(Eds.). Amsterdam: North-Holland, pp.57-70

Suppes, P. and Zinnes, J. (1963). Basic Measurement Theory. En: Luce, R. D., Busch, R. and Galanter, E. (Eds.): *Handbook of Mathematical Psychology*. New York: Wiley.

Thurstone, L. L. (1926). The scoring of individual performance. *Journal of Educational Psychology*, *17*, 446-457.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, *23*, 529-554.

Thurstone, L. L. (1931). Measurement of social attitudes. *Journal of Abnormal and Social Psychology*, *26*, 249-269.

Thurstone, L. L. and Chave, E. J. (1929). *The measurement of attitude*. Chicago: University of Chicago Press.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*. New York: John Wiley and Sons. Inc.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 4, 327-352

Tversky, A. and Gati, I. (1978). Studies of similarity. En E. Rosh and B. Lloyd (Eds.) *On the nature and the principle of formation of categories*. Erlbaun: Hillsdale

Van Maanen, L., Been, P., and Sijtsma, K. (1989). The linear logistic test model and heterogeneity of cognitive strategies. In Roskam, E.E. (Ed.) *Mathematical psychology in progress*. New York: Springer-Verlag.

Van der Linden W. J. and Hambleton R. K. (1997) *Handbook of modern item response theory*. New York: Springer-verlag

Wainer, H., and Messick, S. (1983). *Principals of modern psychological measurement: A Festschrift for Frederic M. Lord.* Hillsdale, NJ: Erlbaum.

Withely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*, 479-494.

Wright, B. D. (1999). Fundamental measurement for Psychology. En Embreston, S. E. and Hershberger, S. L. (Ed.) (1999) *The new rules of measurement*. Mahwah, New Jersey: LEA. pp. 65-104

Wright, B. D. and Linacre, J. M. (1998). *WINSTEPS : A Rasch computer program*. Chicago: MESA Press.

GLOSARIO DE TÉRMINOS

Aplicaciones continuas y continuidad de las operaciones algebraicas: Si son E y E´ dos espacios métricos y son d y d´ las distancias en E y en E´, una aplicación f de E en E´ se dice *continua en un punto* $x_0 \in$ E si, para cada entorno V´de f(x_0) en E´, existe un entorno V de x_0 en E, tal que f (V) ⊂ V´. Se dice que f *es continua en* E si es continua en cada punto de E.

Una aplicación *uniformemente continua* de E en E´ es una aplicación tal que para cada ε> 0 existe un δ> 0 tal que la distancia d(x, y) < δ en E, implica que en E´ sea la distancia d´(f(x), f(y)) < ε.

Una aplicación uniformemente continua es continua, pero el recíproco no es cierto en general.

En la recta real:

a) La aplicación (x, y) → (x+y) de $\mathbb{R} \times \mathbb{R}$ en $\mathbb{R}$ es uniformemente continua.

b) La aplicación (x, y) → xy de $\mathbb{R}$ x $\mathbb{R}$ en $\mathbb{R}$ es continua y para cada a ∈ $\mathbb{R}$ la aplicación x →ax es uniformemente continua.

c) Las aplicaciones (x, y) → sup (x, y); (x, y)→ inf (x, y) son uniformemente continuas.

Base: Una familia de conjuntos abiertos no vacíos se dice que es una *base* para los conjuntos abiertos de un espacio métrico E, si cada conjunto abierto no vacío de E es la reunión de una subfamilia de esa familia de conjuntos abiertos.

Campos y σ-campos: En relación con la medibilidad y la medida hay dos tipos de clases que tienen un papel importante, los campos y los σ-campos.

Un campo es una clase (no vacía) cerrada para todas las operaciones de conjunto finitas. Contiene a $\emptyset$ y a Ω.

Un σ-campo es una clase (no vacía) cerrada para todas las operaciones de conjunto numerables; evidentemente cada σ-campo es un campo.

Clases de equivalencia: La forman todos los pares relacionados mediante la relación binaria R. Llamaremos clase de equivalencia de $x \in A$, a $[x] = \{y \in A; xRy\}$

Complitud: Ver espacios métricos completos

Conceptos topológicos: Los conceptos de entorno, conjuntos abiertos o cerrados, conjuntos densos, punto adherente, adherencia, frontera, etc. son conceptos topológicos, dependen de la topología del espacio que se este considerando. La familia de conjuntos abiertos de un espacio métrico E es lo que frecuentemente se denomina topología de E. Por lo tanto, distancias equivalentes son las que dan lugar a la misma topología.

No son conceptos topológicos las nociones de bola, de esfera, de conjunto acotado, la continuidad uniforme de las funciones, etc.

Las propiedades topológicas de un espacio métrico son invariantes respecto a los homeomorfismos.

Al ser un homeomorfismo una función continua con inversa continua, es una función que de forma biunívoca transforma abiertos en abiertos y, por lo tanto, transforma todas las propiedades topológicas de un espacio en otro.

Conjuntos abiertos: En un espacio métrico E, con distancia d, un *conjunto abierto* es un conjunto A de E en el que para cada $x \in A$, existe un $r > 0$ tal que, en la bola B de radio r, está $B(x; r) \subset A$.

El conjunto vacío es un conjunto abierto
El espacio total E es abierto
Una bola abierta es un conjunto abierto
La unión de una familia de conjuntos abiertos es un abierto
La intersección de un número finito de conjuntos abiertos es un abierto.

Conjuntos cerrados: En un espacio métrico E, un conjunto cerrado es, por definición, el complementario de un conjunto abierto.

El conjunto vacío es un conjunto cerrado.
El espacio total E es un conjunto cerrado.
Una esfera es un conjunto cerrado.
La unión de un número finito de conjuntos cerrados es un cerrado.
La intersección de cada familia de conjuntos cerrados es un conjunto cerrado.

Conjunto cociente: Es el conjunto de todas las clases de equivalencia de A (establecidas mediante R) y se denota por A/R.

Si se define una aplicación sobreyectiva $f : A \rightarrow A/R$. tal que $f(x) = [x]$, esa aplicación se llama *aplicación cociente*.

Es importante señalar que $[x]$ es un elemento del conjunto cociente A/R, y por otra parte, es un subconjunto de A.

Las relaciones de equivalencia lo que hacen es considerar iguales (equivalentes) a todos los elementos relacionados y a su vez, toda clase de equivalencia es un punto en el conjunto cociente.

Continuidad uniforme: Se dice que una función $f : A \rightarrow \mathbb{R}$ es continua en el intervalo $[a, b]$ si para cualquier $\varepsilon > 0$ y para todo $y \in [a, b]$, existe un $\delta > 0$ tal que

$$|x - y| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon$$

Nótese que en este caso δ depende, en general, no sólo de ε sino también de a.

Se dice que una función $f : A \to \mathbb{R}$ es uniformemente continua en un intervalo cerrado [a, b], si para todo $\varepsilon > 0$ existe un $\delta > 0$ tal que si para todos x e y del intervalo [a, b], tales que $|x - y| < \delta$, entonces $|f(x) - f(y)| < \varepsilon$.

Continuidad uniforme en un espacio métrico: Si son A y B dos espacios métricos, diremos que una función $f : A \to B$ es continua uniforme, o uniformemente continua, si para cada $\varepsilon > 0$ existe un $\delta > 0$, tal que , para una distancia d es

$$d(f(x), f(y)) < \varepsilon \quad \text{si} \quad d(x, y) < \delta$$

Observación: El concepto de continuidad en un punto es un concepto local, esto es, dado un ε, el valor del δ elegido depende de ε y además del punto en el espacio métrico. Por lo tanto, hay que elegir un δ que garantice la desigualdad anterior. Toda función uniformemente continua es continua. El reciproco no es cierto.

Distancias equivalentes: Cuando dos distancias d_1, d_2, en un conjunto E, definen sendos espacios métricos E_1, E_2, si la aplicación identidad $x \to x$ de E_1 sobre E_2 es un homeomorfismo, entonces se dice que d_1, d_2 son distancias equivalentes, o topológicamente equivalentes en E. Esto significa que las familias de conjuntos abiertos son las mismas en E_1 y E_2.

Espacios compactos: Un espacio métrico E se llama *compacto* si satisface la siguiente condición, que es el denominado axioma de Borel-Lebesgue: para cada recubrimiento de E mediante conjuntos abiertos (recubrimiento abierto) existe una subfamilia finita de recubrimientos, que es un recubrimiento de E.

Un espacio métrico E se llama *precompacto* si para cada $\varepsilon > 0$, existe un recubrimiento finito de E por medio de conjuntos de diámetro menor que ε.

La definición de *compacidad*, es una noción topológica, en tanto que la de *precompacidad* no lo es.

En la teoría de los espacios métricos, estas nociones son las que sustituyen a la noción de "finitud" en la teoría de conjuntos. En una terminología análoga se podría decir que expresa que el espacio métrico es "aproximadamente finito".

Las tres condiciones siguientes son equivalentes para un espacio métrico:

a) E es compacto
b) cada sucesión infinita de E tiene al menos un valor adherente
c) E es precompacto y completo.

Si es E un espacio métrico, dos de las tres propiedades siguientes, implica la tercera:

a) E es compacto
b) E es discreto, u homeomórfico al espacio discreto
c) E es finito

Las propiedades a) y b) implican la c) y recíprocamente, c) implica a ambas, a a) y a b).

En un espacio métrico compacto E, cada sucesión $\{x_n\}$ que tiene un solo valor adherente a, converge hacia ese valor a.

Toda aplicación continua f de un espacio métrico E que sea compacto, en otro espacio métrico E´, es uniformemente continua.

Espacios conexos: Un espacio métrico E se dice que es *conexo*, si los únicos subconjuntos de ese espacio E que son a la vez abiertos y cerrados son el conjunto vacío $\varnothing$ y el propio conjunto E.

Un espacio métrico E se dice que es *localmente conexo* si, para cada punto $x \in$ E, existe un sistema fundamental de entornos conexos de x.

La recta real $\mathbb{R}$ es un espacio conexo y localmente conexo.

La condición necesaria y suficiente para que un subconjunto A de la recta real $\mathbb{R}$ sea un subconjunto conexo, es que A sea un intervalo, sea o no acotado ese intervalo.

Espacios localmente compactos: Un espacio métrico E se dice que es *localmente compacto* si para cada punto $x \in E$, existe un entorno compacto de x en E.

La recta real $\mathbb{R}$ es localmente compacta pero no es compacta. Esto es una consecuencia inmediata del teorema Borel- Lebesgue.

En un espacio métrico E localmente compacto, cada subespacio abierto y cada subespacio cerrado son localmente compactos.

Espacios métricos completos, sucesiones de Cauchy: Una sucesión de Cauchy en un espacio métrico E, es una sucesión $\{x_n\}$ tal que para cada $\varepsilon > 0$, existe un numero entero n_0 tal que para números r, s tales que $r \geq n_0$ y $s \geq n_0$ implica que $d(x_r, x_s) < \varepsilon$

Se puede afirmar que:

a) Cada sucesión convergente es una sucesión de Cauchy.

b) Si $\{x_n\}$ es una sucesión de Cauchy, cada valor adherente de $\{x_n\}$ es un límite de esa sucesión $\{x_n\}$.

Un espacio métrico E se dice *completo* si cada sucesión de Cauchy en E converge naturalmente hacia un punto de E.

Un subespacio F de un espacio métrico E es completo, entonces el subespacio F es cerrado en E.

Es fácil probar que: *La recta real $\mathbb{R}$ es un espacio métrico completo*

Espacios separables: Un espacio métrico E se dice que es separable si existe en E un conjunto denso numerable o finito.

La recta real $\mathbb{R}$ es separable.

La condición necesaria y suficiente para que un espacio métrico E sea separable, es que exista una base finita o numerable para los conjuntos abiertos de E.

Formas canónicas: Si es G un conjunto de transformaciones sobre cualquier conjunto o espacio S, se dice que dos elementos x, y de S son equivalentes con respecto al grupo G (o también se dice "para el grupo G") si y sólo si existe alguna transformación T en G que aplica x en y. Esta transformación tiene inversa, es T^{-1}, que aplica y en x, luego la relación es simétrica. Además, al tener G la estructura de grupo, se demuestra que también es reflexiva y simétrica; es, por lo tanto, una relación de equivalencia.

Un subconjunto C de S recibe el nombre de clase forma canónica respecto a G, si para todo $x \in S$, es equivalente respecto a G, a uno y sólo a un elemento c en C; este elemento c es entonces la forma canónica de x. Una función f(x) definida para todos los elementos x de S y con valores en otro conjunto, que de ordinario es un conjunto de números, es un *invariante* con respecto a G, si $f(xT) = f(x)$ para todo elemento x de S y toda transformación T de G.

Un conjunto de invariantes $f_1, \ldots, f_n$ es *un conjunto de invariantes* con respecto a G si $f_1(x) = f_1(y), \ldots, f_n(x) = f_n(y)$ implica que x es equivalente a y.

Funciones monótonas: Sea E un conjunto no vacío de la recta ampliada $\overline{\mathbb{R}}$. Una aplicación f de E en $\overline{\mathbb{R}}$ se denomina creciente, creciente en sentido estricto, decreciente, o decreciente en sentido estricto, si la relación $x < y$ en E, implica respectivamente que $f(x) \leq f(y)$, que $f(x) < f(y)$, que $f(x) \geq f(y)$, o que $f(x) > f(y)$. Una función que es creciente o decreciente en sentido estricto, se denomina *monótona* (en sentido estricto).

Si dos funciones f, g son crecientes y esta definida la operación $f + g$, entonces $f + g$ es creciente. Además si ambas son finitas y una de ellas es estrictamente creciente, entonces $f + g$ también es creciente en sentido estricto.

Homeomorfismo: Una aplicación f de un espacio métrico E en un espacio métrico E´ se denomina homeomorfismo si cumple las dos condiciones siguientes:

1) es una biyección
2) f y su aplicación inversa f^{-1} son ambas continuas

A estas aplicaciones también se les llama bicontinuas y la aplicación inversa f^{-1} es también un homeomorfismo de E´en E.

Dos espacios métricos son *homeomorfos* si existe un homeomorfismo entre ellos, y dos espacios homeomorfos a un tercero, son homeomorfos entre si.

Homeomorfismo y clases de equivalencia: La relación de homeomorfismo definida para la clase de todos los espacios topológicos es de equivalencia y el objetivo de la topología es determinar qué espacios pertenecen a la misma clase de equivalencia, en lugar de estudiar cada uno de forma individual. De ahí el gran calado del establecimiento de relaciones homeomórficas, porque no es tan sólo que existe una relación biunívoca entre los espacios, sino que también establece esa relación entre los subconjuntos, que son los abiertos, de las topologías respectivas.

Homomorfismo: Aplicación de un conjunto A sobre otro A´ que conserva las leyes de composición interna existentes en A y en A´. A saber: los puntos o elementos de A´ son imágenes de los de A, conservándose en A´ las propiedades de las operaciones definidas en A: asociatividad, conmutatividad, existencia de elemento neutro, y de elemento simétrico para uno dado. Además, si en A hay definidas dos leyes de las cuales una es distributiva respecto a la otra, las leyes correspondientes en A´, poseen las mismas propiedades.

Por ejemplo, si es f la aplicación homomórfica, y la ley de composición interna en A es la suma, entonces, para elementos cualesquiera a, $b \in A$ se tiene para sus imágenes en A´ que:

$$f(a + b) = f(a) + f(b)$$

Isomorfismo: Es un caso especial de homomorfismo, en el que existe homomorfismo de A sobre A´, y homomorfismo de A´ sobre A.

Monotonía: Ver además funciones monótonas**.**

Una función entre conjuntos ordenados se dice monótona (o isótona) si conserva el orden dado. En matemáticas, estas funciones surgieron en primer lugar en el cálculo, y posteriormente se generalizaron al entorno más abstracto de la teoría del orden. Aunque los conceptos generalmente coinciden, las dos disciplinas han desarrollado una terminología ligeramente diferente; mientras en cálculo se habla de funciones monótonas crecientes (o monótonamente crecientes) y monótonas o monótonamente decrecientes (o simplemente crecientes y decrecientes), en la teoría del orden se usan los términos monótona y antítona, o se habla de funciones que conservan e invierten el orden. Formalmente, si son A y B dos conjuntos, diremos que una función $f : A \rightarrow B$ es monótona si, siempre que para todo x e y sea $x \leq y$, se tiene que $f(x) \leq f(y)$. Si en la definición anterior, la relación $\leq$ se reemplaza por el orden estricto $<$, la función que lo cumpla se dice que es monótonamente creciente o *monótona creciente*, o bien será *monótona decreciente* en sentido estricto si la función invierte el orden, es decir, si para $x < y$, se tiene que $f(x) > f(y)$.

Punto adherente: Un punto de un subconjunto A de E se llama *punto adherente* cuando es un punto $x \in E$ tal que cada entorno de x tiene una intersección no vacía con A. Al conjunto de todos los puntos adherentes de A se denomina la *adherencia* de A. Así, cuando se dice que x no es un punto adherente de A, lo que se indica es que es un punto interior a E - A. Es decir, *la adherencia de un conjunto A es el complementario del exterior de A.*

Relación binaria: Sea A un conjunto de elementos y sean (x, y) un par de elementos del conjunto producto AxA. Una relación de equivalencia R es un subconjunto $R \subset A \times A$, que cumple las propiedades:

1) reflexiva: para todo $x \in A$ se cumple xRx.
2) simétrica: si xRy, entonces yRx.
3) transitiva: si xRy, yRz, entonces xRz; para todo $x,y,z \in A$

Relación binaria conexa: Si es A un conjunto finito de elementos y es R una relación binaria, se dice que esa relación es conexa si está definida para todos los elementos en A.

Subconjuntos densos: En un espacio métrico E, un conjunto A se dice que es *denso con respecto a un conjunto* B, si cada punto de B es un punto adherente de A. Esto es equivalente a decir, que para cada $x \in B$, cada entorno de x contiene puntos de A.

Si A es denso con respecto a B, y B es denso con respecto a C, entonces A es denso con respecto a C.

Un conjunto A denso con respecto a E se denomina *denso por todas partes*, o simplemente *denso* en E; estos conjuntos están caracterizados porque cada conjunto abierto no vacío contiene un punto de A.

El conjunto Q de los números racionales es denso en $\mathbb{R}$.

Teorema de Borel-Lebesgue: Para que un subconjunto de la recta real $\mathbb{R}$ sea relativamente compacto, es necesario y suficiente que sea acotado.

La condición necesaria y suficiente para que un subconjunto A de un espacio métrico E sea relativamente compacto es que cada sucesión de puntos de A tenga un valor adherente en E.

Topología: La topología trata de estudiar la forma intrínseca de un espacio a través de las propiedades que permanecen invariantes por transformaciones continuas. En la recta real, una topología usual es la generada por los intervalos abiertos.

Transformación admisible: Una transformación (generalmente de los valores de una escala) se dice admisible si el sistema numérico obtenido, al sustituir los valores originales por los valores transformados, sigue representando al sistema empírico del que es imagen. Es decir, siempre que conserve la relación de representación entre los sistemas empírico y numérico.

La validez en la medición psicológica quiere ser una obra de referencia para aquellos que se inician en la investigación psicométrica, al intentar que realicen un acercamiento crítico y reflexivo al tratar la problemática de la validez, tanto de las medidas como de las inferencias que se hacen a partir de ellas, cuestionándose el uso de las pruebas que no haya mostrado su legitimidad como instrumento de medida. En esta obra se trata de exponer, por una parte, los conceptos derivados de la teoría general de la medida, incorporando los conocimientos procedentes de la psicología matemática, desarrollados fundamentalmente en la segunda mitad del pasado siglo, para superar el reto de la medida científica en Psicología. Por otra, tratar de conectar todo ello con los avances que simultáneamente se han ido produciendo en la modelización psicométrica para el desarrollo de instrumentos de medida rigurosos y que, a su vez, permitan describir y representar adecuadamente los constructos psicológicos que se intentan medir, paso previo e ineludible para la obtención de instrumentos útiles y válidos.

Aula Abierta

36196AA01A01

UNIVERSIDAD
NACIONAL
DE EDUCACIÓN
A DISTANCIA

ISBN 84-362-5061-3
36196
9 788436 250619