Papeles del Psicólogo, 2010. Vol. 31(1), pp. 57-66
http://www.cop.es/papeles

LAS TEORÍAS DE LOS TESTS: TEORÍA CLÁSICA Y TEORÍA DE RESPUESTA A LOS ÍTEMS

TEST THEORIES: CLASSICAL THEORY AND ITEM RESPONSE THEORY

José Muñiz
Facultad de Psicología. Universidad de Oviedo

Para una interpretación y utilización adecuada de las propiedades psicométricas de los tests es necesario ir más allá del mero cálculo empírico, y conocer los fundamentos en los que se basan esos cálculos. Con el fin de contribuir a esta comprensión más allá del mero manejo superficial de la fórmulas psicométricas, el objetivo fundamental de este trabajo es presentar de una manera no excesivamente técnica y especializada las dos grandes teorías que guían la construcción y análisis de la mayoría de los tests: la Teoría Clásica de los Tests y la Teoría de Respuesta a los Ítems. En primer lugar se hace un apunte histórico sobre los tests, indicando cómo surgen y evolucionan al hilo de los avances técnicos y estadísticos. Tras razonar acerca de la necesidad de utilizar teorías psicométricas para el análisis y construcción de los tests, se expone la lógica que subyace a la Teoría Clásica de los Tests, así como sus dos variantes más granadas, la Teoría de la Generalizabilidad y los Tests Referidos al Criterio. Luego se subrayan las limitaciones más importantes del enfoque clásico y se exponen los fundamentos de la Teoría de Respuesta a los Ítems, dentro de cuyo marco encuentran una solución satisfactoria algunos de los problemas que el enfoque clásico no había sido capaz de resolver de forma satisfactoria. Finalmente se comparan ambos enfoques, y se concluye indicando la necesidad de conocer las teorías de los tests para una mejor comprensión y utilización de los instrumentos de medida.
Palabras clave: *Tests, Teoría Clásica de los Tests, Teoría de Respuesta a los Ítems, Teorías de los tests.*

For a correct interpretation and proper use of the psychometric properties of tests it is necessary to go beyond the mere empirical calculation, and know the grounds on which these calculations are based. To contribute to this understanding beyond the superficial handling of the psychometric formulas, the main goal of this work is to present, in a not technical way, the two most important theories that guide the development and analysis of most tests: Classical Test Theory and Item Response Theory. First, a historic note about tests and testing is made, indicating the evolution of tests according to the technical and statistical advances. The importance of test theories in order to develop and analyse tests is pointed out, and Classical Test Theory, including Generalizability Theory and Criterion Referenced Tests, is presented. After underlining the limitations of the Classical Test Theory approach, Item Response Theory is presented. Within this new framework some of the limitations of the Classical Test Theory find a proper solution. Finally both approaches are compared, emphasizing the importance of test theories for a correct use and interpretation of psychometric properties of the tests.
Key words: *Tests, Classical Test Theory, Item Response Theory, Test theories.*

Los tests constituyen seguramente la tecnología más sofisticada de la que disponen los psicólogos para ejercer su profesión, por eso no es infrecuente que la sociedad identifique a los psicólogos con los tests. Naturalmente, unos psicólogos utilizan los tests más que otros, dependiendo de su campo profesional y de su forma de trabajar. Los tests son muestras de conducta que permiten llevar a cabo inferencias relevantes sobre la conducta de las personas. Bien utilizados son herramientas claves en la profesión del psicólogo. No conviene olvidar que los tests nacen con un afán de objetividad y justicia, para evaluar a las personas por lo que realmente valen, evitando evaluaciones sesgadas por aspectos tales como la cuna, la clase social, la raza, el sexo, las creencias, las cartas de recomendación, y otros sistemas de evaluación subjetivos. Unas veces estos nobles fines se han alcanzado mejor que otras, pero ésa era y sigue siendo la idea central, evaluar a todos por el mismo rasero.

Correspondencia: José Muñiz. *Facultad de Psicología. Universidad de Oviedo. Plaza Feijoo, s/n. 33003 Oviedo. España. E-mail: jmuniz@uniovi.es*

NOTA HISTÓRICA

¿Cuándo aparecen los tests por primera vez en la historia? Suele citarse como el origen remoto de los tests unas pruebas que los emperadores chinos ya hacían allá por el año 3000 antes de Cristo para evaluar la competencia profesional de los oficiales que iban a entrar a su servicio. Otras muchas huellas antiguas pueden rastrarse, pero los tests actuales tienen sus orígenes más cercanos en las pruebas senso-motoras utilizadas por Galton

(1822-1911) en su laboratorio antropométrico. Pero será James McKeen Cattell (1860-1944) el primero inutilizar el término *test mental*, en 1890. Pronto quedó claro (Wissler, 1901) que estos primeros tests senso-motores no eran buenos predictores de las capacidades cognoscitivas de las personas, y Binet y Simon (1905) darán un giro radical al introducir en su nueva escala tareas cognoscitivas para evaluar aspectos como el juicio, la comprensión y el razonamiento. Terman llevó a cabo la revisión de la escala en la Universidad de Stanford, la cual se conoce como la revisión Stanford-Binet (Terman, 1916), utilizando por primera vez el concepto de Cociente Intelectual (CI) para expresar la puntuación de las personas. La idea del CI había sido propuesta originariamente por Stern, dividiendo la Edad mental por la Edad Cronológica y multiplicando el resultado por 100 para evitar decimales.

La escala de Binet abre una tradición de escalas individuales que llega hasta nuestros días. En 1917 los tests reciben otro gran impulso al aparecer los tests colectivos Alfa y Beta a raíz de la necesidad del ejército norteamericano de reclutar rápidamente soldados para la primera guerra mundial. El test Alfa iba dirigido a la población general y el Beta a personas analfabetas o que no dominaban el inglés. Las pruebas tuvieron mucho éxito y terminada la guerra las empresas y otras instituciones adoptaron de forma entusiasta el uso de los tests para distintos menesteres. Comenzaba así una expansión creciente en el uso y creación de tests de todo tipo. La aparición de la técnica del análisis factorial va a suponer un gran avance en la construcción y análisis de los tests, permitiendo la aparición de las baterías de tests, cuyo representante más genuino serían las *Aptitudes Mentales Primarias* (PMA) de Thurstone (Thurstone, 1938; Thurstone y Thurstone, 1941). En España tuvimos la suerte de que uno de los grandes pioneros de la Psicología Española, Mariano Yela, estudiase en Chicago con Thurstone en los años 40, lo que le permitió introducir en nuestro país todos los avances de la época, e impulsar la Psicometría tanto en el mundo académico, como su implementación aplicada, colaborando activamente en el desarrollo de la empresa TEA (Pereña, 2007). La división de la inteligencia en sus distintos factores o dimensiones dio lugar a la aparición de dos grandes líneas de estructuración de las dimensiones cognoscitivas, lo que ha dado en llamarse la escuela inglesa y la escuela americana. En la primera se da más importancia a un factor central de inteligencia general, que coronaría una estructura en la que luego vendrían dos amplias dimensiones, la verbal-educativa y la mecánico-espacial, en las que se articularían otros muchos factores más específicos. El enfoque americano asume una serie de dimensiones no jerarquizadas que compondrían el perfil cognoscitivo, que por ejemplo en el caso del PMA serían: la comprensión verbal, la fluidez verbal, aptitud numérica, aptitud espacial, memoria, rapidez perceptiva y razonamiento general. Ambos enfoques son compatibles, y tienen mucho que ver con la tecnología estadística utilizada, sobre todo el análisis factorial. Toda esta línea de investigaciones psicométricas sobre la inteligencia culmina en la obra magna de Carroll (1993), donde se sintetizan los grandes avances alcanzados. En España trabajos como los de Juan-Espinosa (1997), Colom (1995), o Andrés-Pueyo (1996) recogen y analizan de forma brillante este campo de trabajo.

Pero no sólo se producen avances en el campo de los tests cognoscitivos, también los tests de personalidad se aprovechan de los avances que se producen en la psicometría. Suele citarse la hoja de datos personales utilizada por Woodworth en 1917 para detectar neuróticos graves como el pionero de los tests de personalidad. Por su parte el psiquiatra suizo Rorschach propone en 1921 su test proyectivo de manchas de tinta, al que seguirán otros muchos tests basados en el principio de la proyección, que asume que ante un estímulo ambiguo, la persona evaluada tenderá a producir respuestas que de algún modo reflejan aspectos importantes de su personalidad. El lector interesado en la historia de los tests puede consultar por ejemplo el libro de Anastasi y Urbina (1998), aquí solo tratamos de dar unas pinceladas para entender lo que sigue.

Tras esta larga andadura de unos cien años, uno puede preguntarse, por curiosidad, cuáles son en la actualidad los tests más utilizados por los psicólogos españoles, y si estos difieren de los que utilizan sus colegas europeos. Pues bien, en una encuesta reciente hecha en seis países europeos los tests más utilizados por los psicólogos españoles fueron: 16PF, WISC, WAIS, MMPI, Beck, STAI, Rorschach, Raven, Bender e ISRA. Estos datos son muy similares a los obtenidos en otros países europeos (Muñiz et al., 2001).

En suma, la historia de los tests es una historia exitosa de la que la psicología tiene que sentirse orgullosa, sin olvidar, claro está, que como ocurre con cualquier tecno-

logía de cualquier campo, en ocasiones su utilización por manos inexpertas ha dejado mucho que desear. Es por ello que en la actualidad distintas organizaciones nacionales (Colegio Oficial de Psicólogos, COP) e internacionales (Federación Europea de Asociaciones de Psicólogos, EFPA; Comisión Internacional de Tests, ITC, Asociación Americana de Psicología, APA) desarrollan numerosos proyectos y actividades para potenciar el uso adecuado de los tests (Muñiz, 1997b; Muñiz y Bartram, 2007; Prieto y Muñiz, 2000).

¿POR QUÉ HACEN FALTA TEORÍAS DE LOS TESTS?

Hemos visto en el apartado anterior una breve reseña histórica de cómo han surgido y han ido evolucionando los tests concretos, pero nada hemos dicho acerca de las teorías que posibilitan la construcción de los tests. Así contado podría pensarse que los tests se van sucediendo sin orden ni concierto, pero nada más lejos de la realidad. A la construcción y análisis de los tests subyacen teorías que guían su construcción y que condicionan y tiñen los tests según los avances teóricos y estadísticos de cada momento.

A la vista de ello uno puede preguntarse con toda razón: ¿por qué hacen falta teorías de los tests? O si se quiere de un modo más pragmático, ¿Por qué y para qué tienen los psicólogos en su carrera la asignatura de Psicometría dedicada fundamentalmente a exponer estas teorías? La razón es bien sencilla, los tests son instrumentos de medida sofisticados mediante los cuales los psicólogos llevan a cabo inferencias y toman decisiones sobre aspectos importantes de las personas. Por tanto hay que asegurarse de que esas inferencias son adecuadas y pertinentes, de lo contrario se puede perjudicar notablemente a las personas que acuden a los psicólogos por la razón que sea. Las teorías estadísticas de los tests van a permitir la estimación de las propiedades psicométricas de los tests para de ese modo garantizar que las decisiones tomadas a partir de ellos son las adecuadas. Sin esas teorías no podríamos estimar la fiabilidad y la validez de los tests, lo cual es imprescindible para poder usar los tests de forma rigurosa y científica. Por supuesto, aparte de estas teorías estadísticas sobre los tests, la construcción de una prueba debe de guiarse por un modelo o teoría psicológica sustantiva que dirige su construcción. En el trabajo de Muñiz y Fonseca-Pedrero (2008) pueden consultarse los pasos fundamentales para llevar a cabo la construcción de un test. Para un análisis más detallado del proceso de construcción de un test pueden verse por ejemplo los trabajos de Carretero y Pérez (2005), Downing y Haladyna (2006), Morales, Urosa y Blanco (2003), Muñiz (2000), Schmeiser y Welch (2006), o Wilson (2005).

Hay dos grandes enfoques o teorías a la hora de construir y analizar los tests, son la Teoría Clásica de los Tests (TCT) y el enfoque de la Teoría de Respuesta a los Ítems (TRI). No se trata aquí de llevar a cabo exposiciones detalladas de estas teorías (en español pueden verse, por ejemplo, en Muñiz, 1997a, 2000, 2005), sino de subrayar los aspectos claves, para que así los usuarios de los tests tengan una idea más cabal y comprendan en profundidad el alcance de las propiedades psicométricas de los tests que están utilizando.

TEORÍA CLÁSICA DE LOS TESTS

El enfoque clásico es el predominante en la construcción y análisis de los tests, así, por ejemplo, los diez tests más utilizados por los psicólogos españoles citados en el apartado anterior, todos ellos, sin excepción, han sido desarrollados bajo la óptica clásica. Sólo este dato ya deja bien patente la necesidad de que los profesionales entiendan perfectamente la lógica clásica, sus posibilidades y sus limitaciones.

Antes de entrar en la lógica de la teoría clásica, hay que señalar que hinca sus raíces en los trabajos pioneros de Spearman de principios del siglo XX (Spearman, 1904, 1907, 1913). Lleva por lo tanto unos cien años en el circuito, así que se ha ganado por méritos propios el adjetivo de clásica. A partir de esos años se produce un rápido desarrollo y para 1950 lo esencial ya está hecho, así que Gulliksen (1950) lleva a cabo la síntesis canónica de este enfoque. Más adelante serán Lord y Novick (1968) quienes lleven a cabo una reformulación de la teoría clásica y abran paso al nuevo enfoque de la TRI que veremos luego. Pero veamos lo esencial del enfoque clásico.

MODELO LINEAL CLÁSICO

Según mi experiencia, tras más de treinta años explicando estas cosas a los estudiantes de psicología, lo que más les cuesta entender es para qué, y por qué, se necesita un modelo o teoría para analizar las puntuaciones de los tests. Pero, ¿donde está el problema?, se preguntan, ahí está el test, ahí están las puntuaciones obtenidas por las personas en el test, unas altas, otras bajas, otras

intermedias, así que adelante, asignemos a cada cual su puntuación. Las cosas no son tan sencillas, el psicólogo, como cualquier otro profesional de otro campo, tiene que asegurarse de que el instrumento que utiliza mide con precisión, con poco error. Y eso mismo vale para cualquier instrumento de medida, bien sea un aparato de la policía para medir la velocidad de los vehículos, el metro para medir las distancias, o el surtidor de la gasolinera para medir los litros de gasolina que nos dispensa. Todos esos instrumentos han de estar homologados, requieren algún indicador del grado de precisión con el que miden, máxime los tests, ya que apoyados en ellos se toman decisiones muy importantes para las vidas de las personas. No es difícil estar de acuerdo en esto, pero el problema es que cuando un psicólogo aplica un test a una persona, o a varias, lo que obtiene son las puntuaciones empíricas que esa persona o personas obtienen en el test, pero eso nada nos dice sobre el grado de precisión de esas puntuaciones, no sabemos si esas puntuaciones empíricas obtenidas se corresponden o no con las puntuaciones que verdaderamente le corresponden a esa persona en la prueba. Bien podría ocurrir que las puntuaciones estuviesen, por ejemplo, algo rebajadas debido a que ese día la persona no está en sus mejores condiciones, o porque las condiciones físicas en las que se desarrolló la aplicación de la prueba no eran las más adecuadas, o porque las relaciones establecidas entre los aplicadores de las pruebas y las personas evaluadas dejaron mucho que desear. Los psicólogos, como les ocurre a los que construyen aparatos dispensadores de gasolina, estamos obligados a garantizar que las puntuaciones de nuestros tests sean precisas, tengan poco error, el problema es que esto no se sabe escrutando directamente las puntuaciones que obtienen las personas en los tests, esas puntuaciones vistas así de frente no nos dicen nada acerca de su grado de precisión. Como no lo podemos hacer así de frente, es por lo que tenemos que dar algunos rodeos, es decir, es por lo que tenemos que plantear algunos modelos que subyacen a las puntuaciones a fin de ser capaces de estimar el grado de precisión de éstas. El error está mezclado con la verdadera puntuación, como la sal en el agua del mar, o el polvo con la paja, y para separarlos necesitamos llevar a cabo algunos procesos y ahí es donde entran las teorías o modelos estadísticos. Modelos para esto ha habido muchos, pero uno de los que se ha mostrado más eficaz y parsimonioso es el modelo lineal clásico propuesto originalmente por Spearman. Entender la lógica y funcionamiento del modelo es muy sencillo, lo que ya es algo más latoso, aunque no difícil, es desarrollar los aspectos formales y deducciones del modelo, lo cual constituye el corpus central de la psicometría, pero para eso ya están los psicómetras, alguien tiene que hacerlo.

¿Qué propuso Spearman a principios del siglo XX que ha tenido tanto éxito en la historia de la Psicología? Spearman propone un modelo muy simple, de sentido común, para las puntuaciones de las personas en los tests, y que ha dado en llamarse modelo lineal clásico. Consiste en asumir que la puntuación que una persona obtiene en un test, que denominamos su puntuación empírica, y que suele designarse con la letra X, está formada por dos componentes, por un lado la puntuación verdadera de esa persona en ese test (V), sea la que sea, y por otro un error (e), que puede ser debido a muchas causas que se nos escapan y que no controlamos. Lo dicho puede expresarse formalmente así: $X = V + e$

Ahora bien, si se ha entendido lo dicho, está justificado decir que con esto poco hemos avanzado, pues si una persona saca en un test 70 puntos de puntuación empírica, el modelo no nos permite saber ni cual es su puntuación verdadera ni el error contenido en esa puntuación. Exactamente así es, tenemos un solo dato, la puntuación empírica (X), y dos incógnitas, la puntuación verdadera (V) y el error (e). Desde ese punto de vista no hemos avanzado nada, tenemos, eso sí, un modelo de puntuación que parece sensato y plausible, pero nada más, y nada menos, pues que el modelo sea plausible es todo lo que se puede pedir para empezar. El error cometido al medir alguna variable con un test (e) puede deberse a muchas razones, que pueden estar en la propia persona, en el contexto, o en el test, una clasificación bastante exhaustiva de las fuentes posibles de error puede consultarse en Stanley (1971). Para poder avanzar Spearman añade tres supuestos al modelo y una definición, veamos cuáles son.

El primer supuesto es definir la puntuación verdadera (V) como la esperanza matemática de la puntuación empírica, que formalmente puede escribirse así: $V = E(X)$. Lo que esto significa conceptualmente es que se define la puntuación verdadera de una persona en un test como aquella puntuación que obtendría como media si se le pasase infinitas veces el test. Se trata de una definición teórica, nadie va a pasar infinitas veces un test a nadie, por razones obvias, pero parece plausible pensar que si

esto se hiciese la puntuación media que esa persona sacase en el test sería su verdadera puntuación.

En el segundo supuesto Spearman asume que no existe relación entre la cuantía de las puntuaciones verdaderas de las personas y el tamaño de los errores que afectan a esas puntuaciones. En otras palabras, que el valor de la puntuación verdadera de una persona no tiene nada que ver con el error que afecta esa puntuación, es decir, puede haber puntuaciones verdaderas altas con errores bajos, o altos, no hay conexión entre el tamaño de la puntuación verdadera y el tamaño de los errores. De nuevo se trata de un supuesto en principio razonable, que formalmente puede expresarse así: r (v,e) = 0.

El tercer supuesto establece que los errores de medida de las personas en un test no están relacionados con los errores de medida en otro test distinto. Es decir, no hay ninguna razón para pensar que los errores cometidos en una ocasión vayan a covariar sistemáticamente con los cometidos en otra ocasión. Formalmente este supuesto puede expresarse así: $r(e_j, e_k)=0$.

Estas asunciones parecen razonables y sensatas, pero no se pueden comprobar empíricamente de forma directa, serán las deducciones que luego se hagan a partir de ellas las que permitan confirmarlas o falsearlas. Tras cien años formuladas y con muchos resultados empíricos detrás, bien podemos decir hoy que las ideas de Spearman han sido de gran utilidad para la psicología.

Además del modelo y de estos tres supuestos, se formula una definición de lo que son Tests Paralelos, entendiendo por ello aquellos tests que miden lo mismo exactamente pero con distintos ítems. Las puntuaciones verdaderas de las personas en los tests paralelos serían las mismas, y también serían iguales las varianzas de los errores de medida.

Pues bien, el modelo lineal, junto con los tres supuestos enunciados, y la definición de tests paralelos propuesta, constituyen el cogollo central de la Teoría Clásica de los Tests. Un curso sistemático de Psicometría consiste en llevar a cabo las deducciones correspondientes para a partir de esos ingredientes llegar a las fórmulas que permiten estimar el grado de error que contienen las puntuaciones de los tests, y que se denomina habitualmente Fiabilidad de los Tests, véase al respecto el trabajo de Prieto y Delgado (2010) en este mismo monográfico. También se obtienen otras fórmulas populares de la psicometría, como la de Spearman-Brown, que permite estimar la fiabilidad de un test cuando se aumenta o disminuye su longitud; o las fórmulas de atenuación que permiten estimar el coeficiente de validez de una prueba si se atenúan los errores de medida, tanto de la prueba como del criterio. Por no hablar de la fórmula que permite estimar los cambios en la fiabilidad de un test cuando varía la variabilidad de la muestra en la que se calcula. En suma, el modelo lineal clásico expuesto, junto con los supuestos asumidos y la definición de tests paralelos están a la base de todas las fórmulas clásicas utilizadas habitualmente por los psicólogos que se valen de los tests en su práctica profesional. Alguien podría decir que para usar estas fórmulas no hace falta saber de donde vienen, ni cual es su fundamento, pero tal aserto no es digno de un psicólogo que se respete a si mismo, a su ciencia, y a su profesión.

De modo que cuando los psicólogos manejan sus coeficientes de fiabilidad y validez para indicar a sus clientes o usuarios en general que los tests que utilizan son precisos, tienen poco error de medida, han de saber que esa estimación de la fiabilidad se puede hacer gracias a este sencillo modelo y a los supuestos planteados hace ya más de cien años.

TEORÍA DE LA GENERALIZABILIDAD Y TESTS REFERIDOS AL CRITERIO

Este enfoque clásico ha generado diversas variantes sobre todo en función del tratamiento dado al error de medida. Ha habido numerosos intentos de estimar los distintos componentes del error, tratando de descomponerlo en sus partes. De todos estos intentos el más conocido y sistemático es la Teoría de la Generalizabilidad (TG) propuesta por Cronbach y sus colaboradores (Cronbach, Gleser, Nanda y Rajaratnam, 1972). Se trata de un modelo de uso complejo, que utiliza el análisis de varianza para la mayoría de sus cálculos y estimaciones.

Otro desarrollo psicométrico surgido en el marco clásico ha sido el de los Tests Referidos al Criterio (TRC). Se trata de tests utilizados fundamentalmente en el ámbito educativo y en la evaluación en contextos laborales. Su objetivo es determinar si las personas dominan un criterio concreto o campo de conocimiento, por tanto no pretenden tanto discriminar entre las personas, como la mayoría de los tests psicológicos, sino evaluar en qué grado conocen un campo de conocimiento denominado criterio, de ahí su nombre. Estos tests se desarrollan a partir de la propuesta de Glaser (1963) y han tenido una gran influencia sobre todo en el ámbito educativo.

Los indicadores psicométricos clásicos desarrollados a partir del modelo lineal clásico no se adaptaban bien a la filosofía de construcción de estos nuevos tests, por lo que se ha desarrollado todo un conjunto de tecnología psicométrica específica para calcular la fiabilidad y validez, así como para establecer los puntos de corte que determinan si una persona domina o no el criterio evaluado (Berk, 1984; Cizek, 2001; Educational Measurement, 1994; Muñiz, 2000).

LIMITACIONES DEL ENFOQUE CLÁSICO

Del enfoque de la teoría clásica bien podría decirse que goza de muy buena salud, hay pocas dudas de su utilidad y eficacia, baste decir, por ejemplo, que la gran mayoría de los tests editados en España, prácticamente todos, están desarrollados y analizados dentro de este marco. Ahora bien, si es así, la pregunta obligada es por qué hacen falta otras teorías de los tests, o, en otras palabras, ¿qué problemas de medición no quedaban bien resueltos dentro del marco clásico para que se propongan nuevas teorías? Pues bien, había dos cuestiones básicas que no encontraban buena solución en la teoría clásica y que hacían que la medición psicológica no fuese homologable a la que exhibían otras ciencias empíricas.

Veamos la primera: dentro del marco clásico, las mediciones no resultan invariantes respecto al instrumento utilizado. Se preguntarán con razón qué quiere decir exactamente esa afirmación un tanto críptica. Es muy sencillo, si un psicólogo evalúa la inteligencia de tres personas distintas con un test diferente para cada persona, los resultados no son comparables, no podemos decir en sentido estricto qué persona es más inteligente. Esto es así porque los resultados de los tres tests no están en la misma escala, cada test tiene la suya propia. Esto puede sorprender a los psicólogos usuarios habituales de la teoría clásica, acostumbrados en la práctica a comparar la inteligencia de personas que han sido evaluadas con distintos tests de inteligencia. Para hacerlo se transforman las puntuaciones directas de los tests en otras baremadas, por ejemplo en percentiles, con lo que se considera que se pueden ya comparar, y de hecho así se hace. Este proceder clásico para solventar el problema de la invarianza no es que sea incorrecto, pero, amén de poco elegante científicamente, descansa sobre un pilar muy frágil, a saber, se asume que los grupos normativos en los que se elaboraron los baremos de los distintos tests son equiparables, lo cual es difícil de garantizar en la práctica. Si eso falla la comparación se viene abajo. No hay duda que lo más deseable científicamente sería que los resultados obtenidos al utilizar distintos instrumentos estuviesen en la misma escala, y todo quedaría resuelto de un plumazo, pues bien, por extraño y contra intuitivo que parezca eso es precisamente lo que va a conseguir el enfoque de la TRI. Este nuevo enfoque de la TRI va a suponer un gran avance para la medición psicológica, propiciando un gran desarrollo de nuevos conceptos y herramientas psicométricas.

La segunda gran cuestión no bien resuelta dentro del marco clásico era la ausencia de invarianza de las propiedades de los tests respecto de las personas utilizadas para estimarlas. En otras palabras, propiedades psicométricas importantes de los tests, tales como la dificultad de los ítems, o la fiabilidad del test, estaban en función del tipo de personas utilizadas para calcularlas, lo cual resulta inadmisible desde el punto de vista de una medición rigurosa. Por ejemplo, la dificultad de los ítems, o los coeficientes de fiabilidad dependen en gran medida del tipo de muestra utilizada para calcularlos. Este problema también encontrará una solución adecuada dentro del marco de la TRI.

Aparte de estas dos grandes cuestiones, había otras menores de carácter más técnico a las que la teoría clásica no daba una buena solución. Por ejemplo, cuando se ofrece un coeficiente de fiabilidad de un test en el marco clásico, como el coeficiente alfa de Cronbach (1951), se está presuponiendo que ese test mide con una fiabilidad determinada a todas las personas evaluadas con el test, cuando tenemos evidencia empírica más que suficiente de que los tests no miden con la misma precisión a todas las personas, dependiendo la precisión en gran medida del nivel de la persona en la variable medida. El nuevo marco de la TRI va a solucionar este problema ofreciendo la Función de Información, que permite estimar la fiabilidad de la prueba en función del nivel de la persona en la variable medida.

Además de estas cuestiones centrales, la TRI va a generar toda una tecnología psicométrica nueva que cambiará para siempre la forma de hacer psicometría; véase por ejemplo en este mismo número monográfico el trabajo de Olea, Abad y Barrada (2010). Ahora bien, conviene dejar muy claro que estos nuevos modelos de TRI de ninguna manera invalidan el enfoque clásico, si bien constituyen un excelente complemento que en determina-

das circunstancias dan solución a problemas mal resueltos en el marco clásico. Ambas tecnologías conviven perfectamente en la construcción y análisis de los tests, igual que coches y aviones lo hacen en el transporte, valga la analogía, unos son aconsejables en determinadas situaciones, y otros lo son en otras.

Veamos los conceptos fundamentales sobre los que se apoyan los modelos de TRI.

TEORÍA DE RESPUESTA A LOS ÍTEMS (TRI)

Como se acaba de señalar en el apartado anterior, la TRI va a resolver algunos graves problemas de la medición psicológica que no encontraban una solución adecuada dentro del marco clásico. Ahora bien, para poder hacerlo tiene que pagar el peaje de formular modelos más complejos y menos intuitivos que el modelo clásico, sin que ello suponga que entrañen dificultades especiales. Pero antes de pasar a exponer los fundamentos de estos modelos, vamos a dar unas breves pinceladas de su nacimiento histórico, para así ayudar al lector a ubicarlos en la historia de la psicología. Quienes estén interesados en una descripción detallada de los aspectos históricos pueden consultar por ejemplo el trabajo de Muñiz y Hambleton (1992), titulado medio siglo de teoría de respuesta a los ítems.

RESEÑA HISTÓRICA

En ciencia pocos avances surgen de repente, de la noche a la mañana, sin incubación, lo más habitual es que se produzca un proceso gradual que en un momento determinado cuaja en una nueva línea de trabajo. Y eso es más o menos lo que ha pasado con la TRI, sus primeros atisbos pueden rastrearse en trabajos pioneros de Thurstone allá por los años veinte (Thurstone, 1925), que se continúan en los cuarenta con las aportaciones de autores como Lawley (1943, 1944) o Tucker (1946). Como se puede ver ya en estos años de pleno dominio de la Teoría Clásica se están dando los primeros pasos de los que luego vendría a denominarse TRI. Esos son los orígenes remotos, pero será el gran psicómetra Frederic Lord (1952) quien en su tesis doctoral dirigida por Gulliksen, el gran sintetizador de la Teoría Clásica, ponga los primeros ladrillos firmes de la TRI. Birnbaum en los años cincuenta aporta nuevos avances, pero será el matemático danés Rasch (1960), quien proponga su hoy famoso modelo logístico de un parámetro. Bien podemos tomar esa fecha como el momento de despegue de la TRI, pero nótese que por estas fechas aún nos movemos a nivel meramente teórico y estadístico, muy lejos de las aplicaciones prácticas de estos nuevos modelos. El gran impulso lo darán Lord y Novick (1968) en su famoso libro, en el cual dedican cinco capítulos al tema. A partir de su libro las investigaciones sobre los modelos de TRI dominarán la psicometría, hasta nuestros días. A partir de esa fecha empiezan a aparecer los programas informáticos necesarios para utilizar los modelos de TRI, tales como BICAL y LOGIST en 1976, BILOG en 1984, MULTILOG, 1983, y otros muchos. En 1980 Lord publicará un influyente libro (Lord, 1980) dedicado a las aplicaciones de la TRI. De esas fechas hasta hoy los avances han sido notorios, y podemos decir que en nuestros días la TRI domina el panorama psicométrico. Una introducción a la TRI en español puede consultarse por ejemplo en Muñiz (1997a), en inglés es muy recomendable el libro de

GRÁFICO 1
CURVA CARACTERÍSTICA DE UN ÍTEM (CCI)

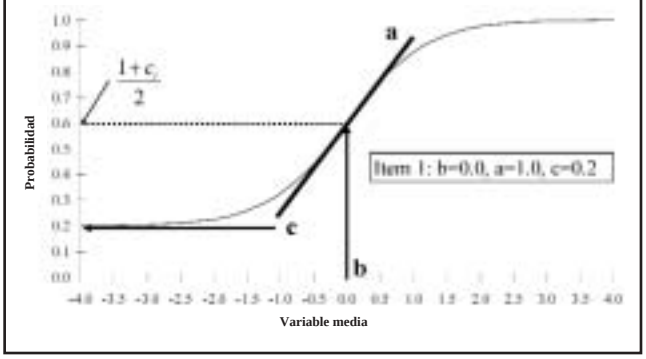
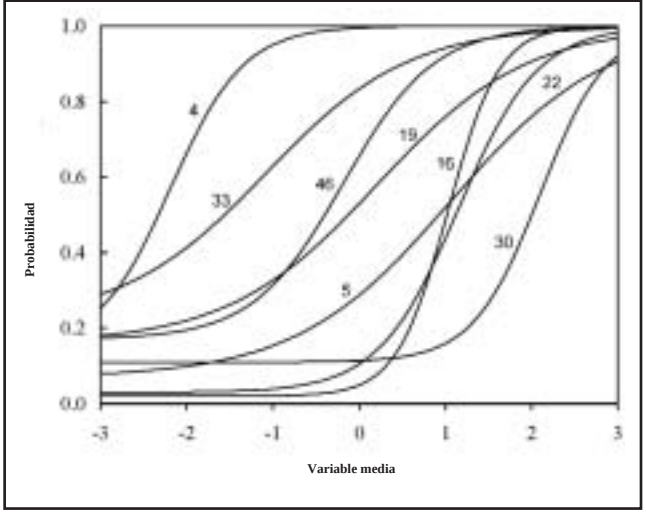


GRÁFICO 2
CURVAS CARACTERÍSTICAS DE OCHO ÍTEMS



Hambleton, Swaminathan y Rogers (1991). Veamos a continuación los supuestos y los modelos de TRI.

SUPUESTOS

Para resolver los problemas citados anteriormente que no encontraban una buena solución dentro del marco clásico, la TRI va a tener que hacer unas asunciones más fuertes y restrictivas que las hechas por la Teoría Clásica. El supuesto clave en los modelos de TRI es que existe una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de acertar estos, denominando a dicha función Curva Característica del Ítem (CCI) (Muñiz, 1997a). Un ejemplo de lo dicho puede verse en el gráfico 1, nótese que al aumentar los valores de la variable medida, denominada θ, aumenta la probabilidad de acertar el ítem $P(\theta)$. Los valores de la variable medida, sea la que sea, se encuentran entre menos infinito y más infinito, mientras que en la teoría clásica los valores dependían de la escala de cada test, yendo desde el valor mínimo obtenible en el test hasta el máximo.

La forma concreta de la CCI viene determinada por el valor que tomen tres parámetros: *a, b* y *c*. Siendo *a* el índice de discriminación del ítem, *b* la dificultad del ítem y *c* la probabilidad que hay de acertar el ítem al azar. Según los parámetros tomen unos valores u otros se generan distintas formas de curvas, como se puede ver en el gráfico 2.

Naturalmente los valores de los parámetros se calculan a partir de los datos obtenidos al aplicar los ítems a una muestra amplia y representativa de personas. Para estos cálculos son necesarios sofisticados programas de ordenador, no en vano los modelos de TRI no se extendieron hasta que se dispuso de ordenadores potentes.

La mayoría de los modelos de TRI, y desde luego los más populares, asumen que los ítems constituyen una sola dimensión, son unidimensionales, por tanto antes de utilizar estos modelos hay que asegurarse de que los datos cumplen esa condición. Esto supone una restricción importante para su uso, pues es bien sabido que muchos de los datos que manejan los psicólogos no son esencialmente unidimensionales, si bien es verdad que los modelos siguen funcionando bastante bien cuando los datos no son estrictamente unidimensionales, es decir son bastante robustos a violaciones moderadas de la unidimensionalidad (Cuesta y Muñiz, 1999).

Un tercer supuesto de los modelos de la TRI es la denominada Independencia Local, que significa que para utilizar estos modelos los ítems han de ser independientes unos de otros, es decir, la respuesta a uno de ellos no puede estar condicionada a la respuesta dada a otros ítems. En realidad si se cumple la unidimensionalidad también se cumple la Independencia Local, por lo que a veces ambos supuestos se tratan conjuntamente.

TABLA 1
DIFERENCIAS ENTRE LA TEORÍA CLÁSICA Y LA TEORÍA DE RESPUESTA A LOS ÍTEMS

Aspectos	Teoría Clásica	Teoría de Respuesta a los Ítems
Modelo	Lineal	No Lineal
Asunciones	Débiles (fáciles de cumplir por los datos)	Fuertes (difíciles de cumplir por los datos)
Invarianza de las mediciones	No	Sí
Invarianza de las propiedades del test	No	Sí
Escala de las puntuaciones	Entre cero y la puntuación máxima en el test	Entre $-\infty$ y $+\infty$
Énfasis	Test	Ítem
Relación Ítem-Test	Sin especificar	Curva Característica del Ítem
Descripción de los ítems	Índices de Dificultad y de Discriminación	Parámetros a, b, c
Errores de medida	Error típico de medida común para toda la muestra	Función de Información (varía según el nivel de aptitud)
Tamaño Muestral	Puede funcionar bien con muestras entre 200 y 500 sujetos aproximadamente	Se recomiendan más de 500 sujetos, aunque depende del modelo

MODELOS

Con los supuestos señalados, según se elija para la Curva Característica de los ítems una función matemática u otra tendremos distintos modelos, por eso se suele hablar de modelos de TRI. Teóricamente habría infinitos posibles modelos, pues funciones matemáticas donde elegir hay de sobra, ahora bien las funciones más utilizadas por razones varias son la función logística y la curva normal. La función logística tiene muchas ventajas sobre la curva normal, pues da resultados similares y sin embargo es mucho más fácil de manejar matemáticamente, así que los tres modelos de TRI más utilizados son los modelos logísticos, que adoptan la función logística como Curva Característica de los ítems. Si sólo se tiene en

cuenta la dificultad de los ítems (parámetro *b*) estamos ante el modelo logístico de un parámetro, o modelo de Rasch, por haber sido propuesto por este autor en 1960 (Rasch, 1960). Si además de la dificultad se tiene en cuenta el índice de discriminación de los ítems (parámetro *a*) estamos ante el modelo logístico de dos parámetros, y si además se añade la probabilidad de acertar el ítem al azar (parámetro *c*), tenemos el modelo logístico de tres parámetros. Este modelo es el más general de los tres, en realidad los otros dos son casos particulares, así cuando el parámetro *c* es cero tenemos el modelo de dos parámetros, y cuando además el parámetro *a* es igual para todos los ítems, se convierte en el modelo de Rasch. Véase a continuación la fórmula del modelo logístico de tres parámetros, donde P(θ) es la probabilidad de acertar el ítem, θ es la puntuación en la variable medida, *a, b* y *c* son los tres parámetros descritos, *e* es la base de los logaritmos neperianos (2,72) y D es una constante que vale 1,7.

$$P(\theta) = c + (1-c)\,[e^{Da(\theta-b)} / (1+e^{Da(\theta-b)})]$$

En la actualidad hay más de cien modelos de TRI, que se utilizan según el tipo de datos manejados, así disponemos de modelos para escalas tipo Likert, para datos dicotómicos, o para datos multidimensionales. Una buena clasificación y revisión de los modelos puede consultarse en el libro de Van der Linden y Hambleton (1997).

COMPARACIÓN DE LA TEORÍA CLÁSICA CON LA TRI

En la tabla 1, tomada de Muñiz (1997a), se sintetizan las diferencias y similitudes entre el enfoque clásico y la TRI.

A MODO DE CONCLUSIÓN

El objetivo de este artículo ha sido el presentar de una manera no técnica a los psicólogos profesionales, lectores de *Papeles del Psicólogo*, las teorías más influyentes en la construcción y análisis de los tests: la Teoría Clásica de los Tests y la Teoría de Respuesta a los Ítems. Espero que estos fundamentos les ayuden a entender e interpretar un poco mejor los datos psicométricos que habitualmente se ofrecen sobre los tests. También sería bueno que ello les animase a refrescar sus conocimientos psicométricos y a profundizar en aspectos nuevos relevantes para su práctica profesional. Todo lo relativo a la medición psicológica ha evolucionado muy rápido en las últimas décadas, produciéndose importantes avances que es necesario seguir de cerca para no quedarse atrás en el ámbito de la evaluación psicológica, pues sin una evaluación precisa y rigurosa no se puede hacer un diagnóstico certero, y sin éste resulta imposible una intervención eficaz.

REFERENCIAS

Anastasi, A., y Urbina, S. (1998). *Los tests psicológicos*. México: Prentice Hall.

Andrés-Pueyo, A. (1996). *Manual de psicología diferencial*. Madrid: McGraw Hill.

Berk, R. A. (Ed.) (1984). *A guide to criterion referenced test construction.* Baltimore, MD: The Johns Hopkins University Press.

Binet, A. y Simon, T. H. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'année Psychologique, 11*, 191-244.

Carretero-Dios, H., y Pérez, C. (2005). Normas para el desarrollo y revisión de estudios instrumentales. *International Journal of Clinical and Health Psychology, 5*, 521-551.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Nueva York: Cambridge University Press.

Cizek, G. J. (Ed.) (2001). *Setting performance standards: Concepts, methods, and perspectives*. Londres: LEA:

Colom, B. R. (1995). *Tests, inteligencia y personalidad*. Madrid: Pirámide.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L.J., Gleser, G., Nanda, H., y Rajaratnam, N. (1972). *The dependability of behavioral measurement: Theory of generalizability for scores and profiles.* Nueva York: Wiley.

Cuesta, M. y Muñiz, J. (1999). Robustness of item response logistic models to violations of the unidimensionality assumption. *Psicothema,* Vol. 11, 175-182

Downing, S. M., y Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Educational Measurement: Issues and Practice (1994). Número monográfico dedicado a los treinta años de tests referidos al criterio. Vol. 13, nº 4.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519-521.

Gulliksen, H. (1950). *Theory of mental tests*. Nueva York: Wiley.

Hambleton, R. K., Swaminathan, H., y Rogers, J. (1991). *Fundamentals of item response theory.* Beverly Hills, CA: Sage.

Juan-Espinosa, M. (1997). *Geografía de la inteligencia humana*. Madrid: Pirámide.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edimburg, 61*, 273-287.

Lawley, D. N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edimburg, 62,* 74-82.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs, nº 7.*

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: LEA.

Lord, F. M., y Novick, M. R. (1968). *Statistical theories of mental test scores*. New York: Addison-Wesley.

Morales, P., Urosa, B., y Blanco, A. B. (2003). *Construcción de escalas de actitudes tipo Likert*. Madrid: La Muralla.

Muñiz, J. (1997a) Introducción a la teoría de respuesta a los ítems. Madrid: Pirámide.

Muñiz, J. (1997b). Aspectos éticos y deontológicos de la evaluación psicológica. En A. Cordero (ed.), *La evaluación psicológica en el año 2000*. Madrid: Tea Ediciones.

Muñiz, J. (2000). *Teoría Clásica de los Tests*. Madrid: Pirámide.

Muñiz, J. (2005). Classical test models. En B. S. Everitt and D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*. Chichester: John Wiley and Sons. (Vol. 1, pp. 278-282).

Muñiz, J., y Bartram, D. (2007). Improving international tests and testing. *European Psychologist, 12*, 206-219.

Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J. R. y Zaal, J. (2001). Testing practices in European countries. *European Journal of Psychological Assessment*, 17(3), 201-211.

Muñiz, J. y Fonseca-Pedrero, E. (2008). Construcción de instrumentos de medida para la evaluación universitaria. *Revista de Investigación en Educación, 5*, 13-25.

Muñiz, J. y Hambleton, R. K. (1992). Medio siglo de teoría de respuesta a los ítems. *Anuario de Psicología,* 52(1), 41-66.

Olea, J., Abad, F.J y Barrada, J.R. (2010). Tests informatizados y otros nuevos tipos de tests. *Papeles del Psicólogo*, *31(1)*, 97-107

Pereña, J. (2007). *Una tea en la psicometría española*. Madrid: Tea Ediciones.

Prieto, G. y Delgado, A. (2010). Fiabilidad y validez. *Papeles del Psicólogo*, *31(1)*, 67-74

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77*, 65-71.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhague: The Danish Institute for Educational Research.

Schmeiser, C. B., y Welch, C. (2006). Test development. En R. L. Brennan (Ed.), *Educational Measurement (4th ed.)* (pp. 307-353). Westport, CT: American Council on Education / Praeger.

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

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

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

Stanley, J. C. (1971). Reliability. En R. L. Thorndike (Ed.), *Educational Measurement.* Washington: American council on Education.

Terman, L. M. (1916). *The measurement of intelligence*. Boston: Houghton Mifflin.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology, 16*, 433-451.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*, nº 1.

Thurstone, L. L. y Thurstone. T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs*, nº 2.

Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1-13.

Van der Linden, W. J. y Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. Nueva York: Springer-Verlag.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wissler, C. (1901). Correlation of mental and physical traits. *Psychological Monographs, 3*, nº 16.