Anuario de Psicología
1992, nº 52, 41-66

Medio siglo de Teoría de Respuesta a los Ítems

José Muñiz
Universidad de Oviedo
Ronald K. Hambleton
University of Massachusetts at Amherst

En el presente trabajo se lleva a cabo una revisión de las principales aportaciones de la Teoría de Respuesta a los Ítems (TRI) desde sus orígenes hasta nuestros días. El trabajo se divide en tres partes: nota histórica, aportaciones principales y conclusiones. En el apartado correspondiente a la nota histórica se señalan en primer lugar los principales problemas que tenía planteados la psicometría clásica antes de la llegada de los modelos de la TRI, y a los que estos modelos aportarán soluciones novedosas, luego se rastrean con cierto detalle los orígenes históricos de la TRI, y finalmente se comentan los años de crecimiento y expansión. Bajo el epígrafe de aportaciones principales se repasan aquellas áreas de la psicometría en las que la TRI ha tenido un mayor impacto: modelos logísticos, función de información, bancos de ítems, tests a medida, tests referidos al criterio, equiparación de puntuaciones y sesgo. En cada una de estas áreas se señalan los avances aportados por la TRI, así como los problemas aún pendientes. Se concluye, finalmente, afirmando que la TRI ha supuesto el mayor avance de los últimos años en la medición psicológica y educativa, utilizándose esta tecnología para la construcción y análisis de la mayoría de los tests actuales; no obstante, problemas como la estimación precisa de los parámetros, la robustez a las violaciones de los supuestos, o los modelos multidimensionales, requerirán investigaciones futuras para soluciones óptimas.
Palabras clave: *Teoría de Respuesta a los Ítems, Modelos Logísticos.*

In this research, a review of the main contributions to Item Response Theory (IRT) since its conception to the present day is carried out. The paper is divided into three parts: historical note, main contributions, and conclusions. In the historical note, the fundamental problems that classical psychometricians were facing before the arrival of IRT are highlighted, and the IRT solutions to these same problems are discussed. Also, special attention is given to those pioneering works representing the historical origins of IRT. And, finally, the years of development of IRT are traced. In the main contributions section, the most important topics are reviewed:

Dirección de los autores: José Muñiz. Facultad de Psicología, Universidad de Oviedo, Aniceto Sela s/n. 33005 Oviedo.

logistic models, information functions, item banks, adaptive testing, criterion-referenced testing, equating, and bias. In each of these areas emphasis is placed in the most remarkable advances due to IRT, as well as the remaining problems. In the final section, it is concluded that IRT models represent the most outstanding contribution in the last 50 years to psychological and educational measurement. However, problems such as parameter estimation, model robustness, issues and methods associated with multidimensional modeling will require solutions, and these areas will, therefore, be the focus of many future investigations.

Key words: *Item Response Theory, Logistic Models.*

Seguramente algunos lectores se sorprenderán de que la Teoría de Respuesta a los Ítems, en adelante TRI, lleve ya medio siglo rondando por los reales de la psicometría, pero, efectivamente, los inicios de lo que finalmente ha dado en llamarse TRI pueden rastrearse entre los años 30 y 40, incluso antes pueden citarse algunas ideas seminales de Thurstone (Thurstone, 1925, 1927, 1928a y b), aunque todo ello lejos de lo que hoy denominamos TRI.

El objetivo central de este trabajo será precisamente seguir los raíles principales por los que ha ido discurriendo la TRI hasta desembocar en su estatus actual, que claramente se puede tildar de hegemónico dentro de la psicometría. Asimismo se tratarán de apuntar los logros alcanzados, las tendencias que se vislumbran para los próximos años, y los problemas que concentran la mayor parte de la actividad investigadora actual.

Nota Histórica

El antes

Al principio fue la Teoría Clásica (TC), un modelo lineal simple (Spearman, 1904) en el que se asume, con mucha enjundia y parsimonia, que las puntuaciones de los sujetos en un test vendrían afectadas por un error aleatorio, debido a causas varias e ignotas, unas dependientes del sujeto, otras del ambiente externo, otras del instrumento de medida, y otras del propio proceso de medición.

Modelo Lineal Clásico: $X = V + e$

Es el modelo de medida del resto de las ciencias empíricas. Para hacer estimaciones acerca de la cuantía de ese error se asume aleatorio y con una distribución normal. Intentos de sistematizar, clasificar y parcelar el error según las posibles fuentes que lo originan (véase Stanley, 1971; Feldt y Brennan, 1989) se han sucedido en psicometría, siendo el más ambicioso y global el proporcionado por la Teoría de la Generalizabilidad (Cronbach y cols., 1972). Pero las complicaciones

introducidas por estos intentos en relación con las ventajas aportadas, cuando alguna, ha determinado que no se hayan erigido en competidores eficaces de la TC.

Para los años 30-40 la psicometría clásica había establecido sus más afamadas fórmulas operativas, tales como la de Spearman-Brown para estimar la fiabilidad de un test en función de su alargamiento,

$$R_{XX'} = nr_{XX'}/[1+(n-1)r_{XX'}]$$

las fórmulas de atenuación (Spearman, 1904) para estimar en qué medida queda atenuada la validez debido a los errores de medida del test y del criterio.

$$R_{XY} = r_{XY}/\sqrt{r_{XX'}r_{YY'}}$$

o las populares KR_{20} y KR_{21} (Kuder y Richardson, 1937) para evaluar la consistencia interna de las pruebas, luego subsumidas como casos particulares del coeficiente alfa,

$$KR_{20} = [n/(n-1)]\,[1-\Sigma p_i q_i/\sigma_X^2]$$

$$KR_{21} = [n/(n-1)]\,[1-(\bar{X}-\bar{X}^2/n)/\sigma_X^2]$$

sólo por citar algunas de las más conocidas y usadas. Todo funcionaba bien, al menos aparentemente, y Guilford (1936) trató de reunir en su manual *Psychometric Methods* el saber psicométrico de la época, tanto en lo referente a escalamiento de sujetos (Teoría de los tests), como al escalamiento de los estímulos (Métodos psicofísicos). Ambos campos eran por entonces ya tan complejos que nunca más nadie intentó exponerlos en un solo texto. Puede decirse que a partir de ahí psicometría va a identificarse con Teoría de los Tests, tratándose los métodos psicofísicos bajo el epígrafe de escalamiento (*scaling*). Ambos campos fueron revisados conjuntamente por última vez para el *Annual Review* por Tucker (1963); en 1965 Ekman y Sjoberg hacen por separado la revisión correspondiente a *scaling*, y en 1967 Keats lleva a cabo la correspondiente a Teoría de los Tests, que aparecía como entrada específica por primera vez en el *Annual Review* desde su fundación en 1950. Como casi toda partición, ésta tiene algo de arbitrario, pues la mayoría de los modelos podrían generalizarse tanto a estímulos como a sujetos, pero también existen bastantes problemas específicos que justifican la división. Ambas ramas tendrán sus síntesis más clásicas en los textos de Gulliksen (1950) y Torgerson (1958) respectivamente. Para un intento de unificación formal véase Mosier (1940, 1941).

Desde un punto de vista llamémoslo tecnológico, la TC tenía planteados a estas alturas dos problemas fundamentales. Uno, ya citado, relativo a las fuentes de error y su operativización, y cuyas alternativas se movían también dentro del marco clásico, si bien añadiendo algunas complejidades para dar cuenta de las distintas fuentes. Podría decirse que este problema como tal no es específico de la medición psicológica, pues en mayor o menor grado afecta a las mediciones en todas las ciencias empíricas, en menor grado ciertamente en el caso de las ciencias

duras, en las que las fuentes de error suelen ser más restringidas y estar bien identificadas.

Sin embargo, la TC sí tenía planteado un problema muy específico cuya solución adecuada no podía hallarse dentro del propio marco de la TC. Ello no quiere decir que cara a la práctica empírica no se hubiesen diseñado algunos remedios estadísticos razonables. Era *el problema de la invarianza de las mediciones y de las propiedades de los instrumentos de medida.* Es decir, si dos tests distintos miden una misma variable, por ejemplo la inteligencia, y los aplicamos a dos sujetos diferentes no podemos saber a ciencia cierta cuál de ellos es más inteligente, pues los resultados no están en la misma escala. La solución adoptada por la TC era expresar las puntuaciones de forma relativa, en función de un grupo normativo. Esta solución, aunque sensata, y útil en la práctica, no constituye una solución técnicamente aceptable, pues si se aspira a una medición rigurosa y científica resulta difícil justificar que las mediciones estén en función del instrumento utilizado[1]. Algo así como si la longitud de un objeto dependiese del tipo de regla utilizado para medirlo. Este problema no tenía solución dentro del marco clásico, y necesariamente se terminaba definiendo la variable medida por el instrumento utilizado para medirla. La otra invarianza reclamada, era la de las propiedades de los instrumentos de medida respecto de los objetos medidos, a saber, la mayoría de las propiedades de un test, piénsese, por ejemplo, en su dificultad, dependían del tipo de sujetos utilizado para calcularla, lo cual es inadmisible a todas luces para un instrumento de medida. Mutatis mutandis, imagínese por un momento que las cualidades métricas de una balanza dependiesen de los objetos pesados. La exigencia de dar una salida adecuada a esos problemas se acentuaba a medida que se iba generalizando el uso de los tests. Así, por ejemplo, si una institución realiza al año cuatro exámenes de admisión y desea que los sujetos tengan las mismas oportunidades en las cuatro ocasiones, es decir, que los tests sean igual de difíciles en los cuatro casos, se encuentra con la prácticamente imposible tarea de construir cuatro tests paralelos, y todo porque las mediciones no son invariantes respecto de los tests utilizados.

Serán estos dos problemas conectados con la invarianza los que encontrarán una solución adecuada dentro del marco de la TRI. Además, la TRI proporcionará todo un conjunto de avances tecnológicos para la construcción y análisis de los tests que cambiarán radicalmente la forma de hacer psicometría. Véase, por ejemplo, Van der Linden (1986) y Hambleton (1986) para un análisis más detallado de los cambios de enfoque de la medición en los últimos años.

Tal vez convenga dejar claro desde el principio, que los avances aportados

1. In any consideration of the nature of the metric provided by raw scores on a mental test, one is likely to be faced with the fact that the raw score units of measurement cannot ordinarily be considered as «equal». If we administer two tests of the same trait or ability, the two tests having different distributions of item «difficulty», to the same group of examinees, we will obtain two different shapes of raw score distributions from the two tests... Since there is no reason to prefer one of these distributions over the other, and since the two distributions cannot both simultaneously represent the shape of the distribution of the trait or ability in the group tested, we conclude that neither distribution gives a true representation of the shape of the distribution of the trait or ability in the group tested and that the raw score scale does not provide equal units of measurement in the case of either tests. In considering such matters, one is usually involved, either implicitly or explicitly, in the assumption that underlying the raw score of the test there is a trait or ability that it is desired to «measure» (Lord, 1953a, p. 517).

por la TRI van a ser fundamentalmente de carácter tecnológico. Es decir, la TRI y metodologías derivadas van a permitir una construcción y análisis mucho más potente de los tests, pero los problemas de tipo teórico-fundante de la medición psicológica no le son ajenos, y todo lo dicho y escrito sobre ellos sigue siendo aplicable a la TRI. Nos referimos a problemas tales como los del estatuts de los rasgos psicológicos, estabilidad, reificación, circularidad, etc. (Mischel, 1968; Carson, 1969; Vale y Vale, 1969). O problemas relacionados con la Teoría de la Medición, tales como la transitividad de las medidas psicológicas, representación, unicidad, significación, cero absoluto de las escalas psicológicas, etc. (Krantz, y cols., 1971; Luce y Narens, 1986; Michell, 1990; Narens, 1985; Narens y Luce, 1986; Pfanzagl, 1968; Roberts, 1979; Savage y Ehrlich, 1990). Por no citar los problemas relativos a la validez en sus distintas acepciones (Anastasi, 1986; Linn, 1990; Messick, 1989; Wainer y Braun, 1988). Ese tipo de problemas siempre estarán ahí sea cual sea la tecnología psicométrica utilizada, pues provienen de la naturaleza de lo medido, de lo psicológico, no conviene engañarse.

Los orígenes

Los primeros trabajos tentativos que ahora retrospectivamente pueden ser vistos como el germen de lo que se ha dado en llamar posteriormente TRI, pero que entonces ni se vislumbraba donde desembocarían, se deben una vez más a Thurstone (Thurstone, 1925, 1927, 1928a y b; Thurstone y Ackerson, 1929). En especial el trabajo de 1925 podría considerarse como un claro antecedente de las Curvas Características de los Ítems, cuando Thurstone en su figura 5, p. 444, presenta una serie de curvas conectando la edad de los sujetos con la proporción de aciertos en cada ítem, tomados del test de Binet. Tucker (1987) en su revisión de los métodos clásicos de análisis de ítems señala también este trabajo como uno de los pioneros, y se atribuye el haber acuñado por primera vez hacia 1945 el término «Curva Característica del Ítem»[2], acuñación que reconoce Lord (1952, p. 5). Dicho sea de paso, la figura de Thurstone cobra cada día mayor estatura psicométrica, si cabe, representando esa difícil unión entre la relevancia sustantiva de los problemas psicológicos abordados y su formulación matemática rigurosa. Por ejemplo, en relación con el problema de la invarianza de los instrumentos respecto de los objetos medidos antes señalado, sus palabras resultan paradigmáticas ya en 1928: «... un instrumento de medida no debe venir afectado por los objetos medidos... sus mediciones deben ser independientes de los objetos medidos» (Thurstone, 1928b, p. 547).

En estos primeros atisbos, además de los propios Binet y Simon (1905a y b, 1908), cuyos gráficos de la evolución de los niños según la edad pueden ser considerados como una primera aproximación a curvas características rudimentarias, hay que citar también el trabajo de Richardson (1936), que puede considerarse seguramente como el primer intento de ajustar la ojiva normal a las res-

2. About 1945 I became interested, for theoretical purposes, in what I called Item Characteristic Curves (Tucker, 1987, p. 2).

puestas a los ítems. Sus consejos acerca de la necesidad de controlar la dificultad de los ítems en función de los objetivos perseguidos por el test (p. 49) representan una formulación verbal anticipada de lo que luego habría de permitir realizar formalmente la Función de Información. Ferguson (1942) también se acerca, vía los métodos psicofísicos, al planteamiento de las curvas características de los ítems. El paralelismo de tratar las proporciones de aciertos en los ítems frente a los valores globales en el test, en los mismos términos que lo venían haciendo los métodos psicofísicos para la determinación de los umbrales será una característica común en estos comienzos. El propio Ferguson (1942) señala explícitamente que en los últimos años se da una tendencia creciente entre los psicómetras a acercar sus métodos a los de la psicofísica (p. 19). Nada más natural que a la hora de determinar los parámetros de los ítems los teóricos de los tests acudiesen a los métodos psicofísicos clásicos, en concreto al de los estímulos constantes, pues tenían el mismo problema que éstos para determinar el umbral absoluto, aquel valor en el eje de abscisas detectado el 50 % de las veces, para lo cual se estaba utilizando la función psicométrica bajo la hipótesis phi-gamma (Muñiz, 1991), conceptualmente equivalente al parámetro b (dificultad), valor de Θ cuando $P(\Theta)=0.50$, supuesto no aciertos al azar.

Lawley (1943, 1944) lleva a cabo una aproximación más sistemática para modelos muy restrictivos, y Tucker (1946) también utiliza la curva normal como rudimento de CCI. Suele atribuirse a Lazarsfeld (1950) la paternidad del término «rasgo latente», que será el nombre que tomarán en principio los modelos, aunque posteriormente se haya generalizado el de TRI, pues refleja mejor su funcionamiento real, basado en los ítems, y permite distinguirlos de otras modelizaciones que también utilizan el término «latente», como el análisis factorial, ecuaciones estructurales o análisis multidimensional (Hambleton y Swaminathan, 1985), amén de evitar las poco favorables connotaciones de «lo latente» en la teorización psicológica. En su reciente revisión, Goldstein y Wood (1989) proponen que el término «Teoría» se cambie por «Modelos», ya que más que teorías psicológicas explicativas lo que se hace es modelizar las respuestas a los ítems.

Con estos orígenes aún remotos citados, el nacimiento formal podría ubicarse en los trabajos de Lord (1952, 1953a y b), que representan la semilla de la que saldrán los frutos de la TRI actual. El trabajo de Lord (1952), «A theory of test scores», publicado en el *Psychometric Monographs nº 7*, es el resultado de su Tesis Doctoral, dirigida por Gulliksen y asesorado por Tucker, la nata de la época. Representa con sus propios trabajos de 1953 (Lord, 1953a y b) la formulación más sistemática de los principales conceptos de la TRI, a partir de los cuales surgirán los desarrollos posteriores. Si hubiera que ubicar puntualmente en algún lado los orígenes genuinos de la TRI lo haríamos en estos trabajos de Lord, especialmente en el de 1952. La nueva teoría formulada marcará un nuevo rumbo en las investigaciones psicométricas, si bien, como el propio Lord indica, las conclusiones obtenidas no contradicen en general los grandes logros de la Teoría Clásica[3].

3. The present theory of test scores starts with assumptions designed to fit certain testing situations, and proceeds to investigate the shapes of the frequency distributions of test scores, of true scores, and of errors of measurement, and further, the relation of these variables to the «ability» involved in taking the test. The conclusions reached do not in

A partir de entonces surge un nuevo modo de hacer psicometría, aunque todavía se está muy lejos, faltan 30 años, para que los modelos de TRI se impongan claramente en el mercado psicométrico. Birnbaum (1957, 1958a y b) da otro gran empujón en el área, sustituyendo los modelos de ojiva normal de Lord por los logísticos, más tratables matemáticamente, y generando los desarrollos matemáticos necesarios para su posible y futuro uso aplicado.

En 1960 el danés George Rasch publica su famoso libro en el que expone con detalle el modelo logístico de un parámetro, utilizando material de tests de aptitudes, y que estaba llamado a convertirse en el más popular y utilizado en la práctica.

Rasch es completamente consciente de que su trabajo supone un cambio radical en el enfoque psicométrico, y en la introducción expone con claridad cómo su modelo viene a resolver los problemas de invarianza ya mencionados[4].

Nótese que hasta ahora nada de lo hecho se traduce en una aplicabilidad directa y generalizada de los modelos por parte de los posibles usuarios, nos movemos a nivel teórico-matemático. El impulso más potente llegará sin duda con la publicación en 1968 del libro de Lord y Novick en el que se dedican cinco capítulos al tema, cuatro de ellos escritos por Birnbaum.

Llegados ahí, 1968, puede decirse que el grueso del corpus general está escrito, y los primeros modelos formulados, pero la implantación y progreso será lenta y laboriosa, debido a la complejidad matemática de los modelos, a la *ausencia de programas de ordenador* disponibles para analizar los datos según los nuevos modelos, y al escepticismo general acerca de las ventajas de esta nueva línea de investigación (Hambleton y Swaminathan, 1985, p. 7). Bock y Wood (1971) incluyen por primera vez en las revisiones para el *Annual Review* un apartado dedicado a la entonces denominada Teoría del Rasgo Latente, donde exponen con claridad las ventajas de los nuevos modelos y la literatura hasta entonces sobre el tema, con especial hincapié en el libro de Lord y Novick (1968) y el de Rasch (1960), como no podía ser menos.

Crecimiento y expansión

Entre la publicación del libro de Lord y Novick (1968) y la aparición de los principales programas de ordenador, BICAL (Wright y Mead, 1976; Wright,

general contradict the basic formulas already firmly established in mental test theory, such as the Spearman-Brown formula and the formula for correction for attenuation. A number of new conclusions are reached, however. Some of them are at variance with certain commonly held conceptions; for example, it is found that the regression of test score and of true score on ability is in general necessarily curvilinear and that errors of measurement have a binomial distribution that is not independent of true score (Lord, 1952, p. 4).

4. In the present work a new approach to test-psychology is attempted. Traditionally the properties of a test are defined in terms of variations within some specified population. In practice such populations may be reasonably selected in various ways and accordingly the properties referred to, e.g. the reliability coefficient, are not specific to the test itself, they may vary, even strongly, with the population chosen. In the following chapters we are going to deal mainly with three different types of tests and for each of them we shall develop a probabilistic model, in the application of which the role of the population can be abolished... Each model implies two types of parameters, a «difficulty» for each test (or item) and an «ability» for each person. The «difficulties» of the tests of course have to be estimated from the body of data available, i.e. the results in two or more tests (or items) for each person of a certain collection. This collection,

Mead y Bell, 1979), LOGIST (Wood, Wingersky y Lord, 1976; Winkersky, Barton y Lord, 1982), BILOG (Mislevy y Bock, 1983), MULTILOG (Thissen, 1983), MICROCAT (Assessment Systems Corporation, 1988), NOHARM (Fraser, 1981), ANCILLES y OGIVA (Urry, 1977), fundamentales para la utilización de los modelos, transcurre una década de rápido crecimiento de la literatura y los avances en TRI, y se empiezan a vislumbrar con claridad las posibilidades reales de la aplicación práctica de los modelos (Lord, 1977; *Journal of Educational Measurement*, 1977; Wright, 1977a y b; Hambleton y cols., 1978; Hambleton, 1979; Wright y Stone, 1979). Especial mención por su militancia en pro de la TRI merece el profesor Benjamin Wright y su grupo de Chicago. Su conferencia invitada de 1967 (Wright, 1968) en el Educational Testing Service (ETS) de Princeton, espoleó a los grandes constructores de tests que allí moran al uso de los nuevos modelos.

Pero será la década de los 80-90 la que supondrá la verdadera expansión y afianzamiento de la TRI y su masivo predominio en psicometría. El punto de inflexión puede ubicarse en otro libro, como no, de Frederic Lord publicado en 1980, y sintomáticamente titulado *Applications of Item Response Theory to Practical Testing Problems*, pues, efectivamente, las aplicaciones habían llegado. En este excelente libro, hito bibliográfico para la TRI, Lord recoge tanto los desarrollos teóricos como las aplicaciones de la TRI desarrolladas hasta entonces.

A partir de esas fechas los trabajos sobre TRI se multiplicarán y las revistas del área y los congresos son invadidos por trabajos relacionados con la TRI. En 1982 la revista *Applied Psychological Measurement* dedica un número monográfico al tema, y aparecen toda una serie de textos comprensivos que cubren los distintos aspectos que se han ido desarrollando durante los años anteriores, y entre los que merecen destacarse los de Hulin, Drasgow y Parsons (1983), Hambleton y Swaminathan (1985), Baker (1985), Andrich (1988), Linn (1989) y Hambleton, Swaminathan y Rogers (1991), entre otros. Para una introducción en castellano véase Muñiz (1990). Una bibliografía exhaustiva clasificada de la TRI puede consultarse en Hambleton (1990a). Excelentes revisiones de la literatura las llevan a cabo para el *Annual Review*, Weiss y Davison (1981) y Traub y Lam (1985), contribuyendo notablemente con ellas a la difusión de los modelos, véase también Jones Appelbaum (1989).

Principales Aportaciones

Modelos Logísticos

Esta expansión de la TRI ha llevado consigo, por un lado, la aplicación de los modelos a las más diversas áreas sustantivas de la Psicología y la Educa-

however, is not taken to be a sample from any «population». On the contrary, the estimation procedure may be so conducted that the personal parameters —the «abilities»— and their possible distribution are eliminated. Similarly the «ability» of each person has to be estimated from the results of the tests applied to him, but the estimation procedure yields a result that is independent of which particular set of tests (or items) has been employed. Thus the parameters stand for intrinsic properties of the tests (or items) and the persons, and they can be estimated accordingly... Our concepts are more akin to the psychophysical measurements in so far as these are concerned with individuals, each observed several times (Rasch, 1960, pp. 3-4).

ción (ver Thissen y Steinberg, 1988), y por otro, a la *multiplicación de los modelos.* Éstos pueden clasificarse atendiendo a distintos criterios, según el número de parámetros a estimar, según el tipo de respuesta que reclamen los ítems, según el número de dimensiones que se asuman, etc. Si bien es cierto que *los más usados e investigados son con diferencia los logísticos de 1, 2 y 3 parámetros* entre otras razones por los programas disponibles para su uso (LOGIST, BICAL, BILOG, MICROCAT). Thissen y Steinberg (1986) ofrecen una clasificación interesante en cinco categorías: *Binary* (Lawley, 1943; Tucker, 1946; Lord, 1952; Rasch, 1960; Birnbaum, 1968; Winsberg, Thissen y Wainer, 1983), *Difference* (Samejima, 1969), *Divide-by-Total* (Masters, 1982; Andrich, 1978; Masters y Wright, 1984; Bock, 1972), *Left-Side-Added* (Birnbaum, 1968; Choppin, 1983), *Left-Side-Added Divide-by-Total* (Samejima, 1972; Sympson, 1983; Thissen y Steinberg, 1984). También Goldstein y Wood (1989) proponen una ordenación de los modelos dentro de un marco general de modelización lineal, y McDonald (1982) presenta excelentes criterios para la clasificación, ordenación y generación de nuevos modelos. Como señala Hambleton (1989), no existe un límite para el número de modelos que pueden ser generados dentro del marco de la TRI, subrayando por su parte, que a los clásicos logísticos de 1, 2, 3 y 4 parámetros, y sus correspondientes de ojiva normal, se han ido sumando en la literatura actual modelos unidimensionales para respuestas politómicas (ordenadas, no ordenadas, o continuas), extensiones del modelo de un parámetro que permiten la incorporación de componentes cognitivos, modelos multidimensionales, así como modelos en los que la unidad de análisis pasa a ser el grupo o clase en vez del sujeto individual. Para una descripción de estos modelos véase la obra citada de Hambleton (1989), limitándonos aquí a comentar las características más destacadas de los modelos logísticos unidimensionales de 1, 2 y 3 parámetros.

Supuestos

Los modelos logísticos de 1, 2 y 3 parámetros asumen la Función Logística como función matemática que relaciona los valores de la variable medida (Θ) con la probabilidad de acertar el ítem $P(\Theta)$, denominada Curva Característica del Ítem (CCI). La suma de las CCI generará la Curva Característica del Test.

Según se contemple sólo el parámetro de dificultad (b), dificultad y discriminación (a), o se añada a estos dos un tercero de aciertos al azar (c), se tendrán respectivamente los tres modelos logísticos de 1, 2 o 3 parámetros. La función que conecta los valores b de los ítems con la proporción de aciertos en esos ítems de cada sujeto determina la Curva Característica del Sujeto, de gran interés teórico cuando se comparan la curva empírica generada por el sujeto con la teórica que cabría esperar según su competencia y el modelo ajustado. Probablemente sea éste uno de los aspectos que no ha recibido toda la atención que se merece, pero que a buen seguro tenderá a desarrollarse en el futuro, pues constituye una fuente potencial de interesantes hipótesis acerca del comportamiento individual y estrategias de los sujetos ante los ítems, permitiendo establecer un puente entre el sujeto individual y el modelo general estimado. Algunos trabajos que apuntan

en esa dirección son los de Drasgow (1982), Drasgow y cols. (1987), Levine y Drasgow (1982) y Levine y Rubin (1979).

Modelo de Rasch: $P(\Theta)=e^{D(\Theta-b)}/[1+e^{D(\Theta-b)}]$
Modelo Logístico de 2 parámetros: $P(\Theta)=e^{Da(\Theta-b)}/[1+e^{Da(\Theta-b)}]$
Modelo Logístico de 3 parámetros: $P(\Theta)=c+(1-c)[e^{Da(\Theta-b)}]/[1+e^{Da(\Theta-b)}]$

Asimismo, los modelos asumen implícitamente en su formulación que los ítems que miden la variable Θ constituyen una sola dimensión, son unidimensionales. En otras palabras, se asume que acertar o fallar un ítem sólo depende de Θ. Esta unidimensionalidad implica también la existencia de *Independencia Local*, es decir, la no dependencia de un ítem del resto de los ítems, pues de lo contrario se caería en la contradicción de que la respuesta a un ítem no depende sólo de Θ como se afirma con la unidimensionalidad. En el caso de los modelos multidimensionales la CCI recibe el nombre más general de Función Característica, pues ya no es una curva en el plano, sino una función, la que sea, en un hiperplano. Aunque la estimación de sus parámetros aún tiene algunos problemas técnicos, se han propuesto diversos modelos multidimensionales (Samejima, 1974; Whitely, 1980; Bock y Aitkin, 1981; Thissen y Steinberg, 1984), constituyendo una de las líneas de investigación futura más relevante, pues la mayoría de las variables psicoeducativas de interés lo más probable es que tengan una naturaleza compleja y multidimensional, para cuya evaluación se requieren estos modelos. Existe también un amplia corriente de opinión entre los especialistas de la medición, mayormente europeos, más proclive a que los instrumentos de medida sean unidimensionales, y si hay que evaluar varias dimensiones utilizar un instrumento unidimensional para cada una de ellas. Los instrumentos de medida se acercarían de este modo más a los ideales de medida de las ciencias experimentales. Parece como si se reprodujesen aquí, salvando las distancias, las posturas relativas a los modelos de análisis factorial de la escuela inglesa (Europa) y de la americana. La dicotomía no es baladí, y puede determinar la dirección de los esfuerzos investigadores futuros. Si en ambos casos los modelos resultantes funcionasen perfectamente las soluciones serían equivalentes, quedando todo en una declaración académica de principios filosóficos. Por un lado, los problemas matemáticos implicados en la estimación de los parámetros de los modelos multidimensionales son notablemente más complicados, pero, por otro, es dudosa la posibilidad de construir instrumentos para cada dimensión, cuando éstas, como suele ser el caso, están altamente interconectadas. Quizá una tercera vía de modelos multidimensionales restringidos a un número bajo de dimensiones represente una salida más viable en la práctica.

La propia determinación de la dimensionalidad ya constituye un verdadero problema no bien resuelto aún. Hattie (1985) cita hasta 87 índices utilizados en la literatura como indicadores de unidimensionalidad. Trabajos acerca de algunos de estos índices pueden verse en Berger y Knol (1990), Hattie (1984), Hambleton y Rovinelli (1986), McDonald (1981), Stout (1987), o Zwick (1987). Diferentes investigaciones sugieren que los modelos logísticos resultan bastante robustos a violaciones no exageradas de la unidimensionalidad (Ansley y Forsyth, 1985;

Drasgow y Parsons, 1983; Folk y Green, 1989; Greaud, 1988; Harrison, 1986; Reckase, 1979; Reckase y cols., 1988; Way y cols., 1988; Yen, 1984), pero son necesarios más datos para determinar con precisión la relevancia de las violaciones.

Estimación de los parámetros

Como en cualquier otro modelo, un aspecto vital han sido las investigaciones acerca de la estimación de los parámetros a las que se ha dedicado notable esfuerzo y es previsbile que se siga haciendo en el futuro. Los avances en lo que a la estimación de los parámetros se refiere han sido enormes en estos últimos años (Drasgow, 1989; Traub y Lam, 1985), pudiendo afirmarse que en líneas generales el problema de la estimación en los modelos logísticos está bien resuelto (Thissen, 1982). El método de estimación más utilizado ha sido el de Máxima Verosimilitud, aunque son cada día más frecuentes las aproximaciones bayesianas (Lord, 1986; Mislevy, 1986; Swaminathan, 1983; Swaminathan y Gifford, 1982, 1985, 1986; Tsutakawa y Lin, 1986). Exposiciones pormenorizadas acerca de la estimación de los parámetros pueden consultarse en Baker (1987), Birnbaum (1968), Hambleton y Swaminathan (1985), Lord (1980) o Swaminathan (1983). Como ya se ha señalado los programas más utilizados siguen siendo LOGIST, BILOG, BICAL y MICROCAT, todos ellos ofrecen como salida fundamental los valores estimados de los parámetros de cada ítem y el valor Θ de cada sujeto. Yen (1987) y Mislevy y Stocking (1989) llevan a cabo estudios técnicos pormenorizados comparando la eficacia y precisión de los programas LOGIST y BILOG. Mislevy y Stocking (1989, pp. 73-74) concluyen al respecto que para las aplicaciones en las que es recomendable LOGIST (tests largos y grandes muestras, así como cuando algunos ítems se omiten o no llegan a contestarse), ambos programas generan estimaciones parecidas, por lo que atendiendo al coste podría preferirse LOGIST. Sin embargo, para los usuarios con tests cortos y muestras no muy numerosas, BILOG sería más recomendable. De todas formas los programas están en continua revisión y mejora, por lo que estas comparaciones pueden quedarse anticuadas rápidamente.

La robustez de las estimaciones ante violaciones de los supuestos ha sido ampliamente investigada, con buenas revisiones del estado de la cuestión en Baker (1987) y Lord (1986). El parámetro c (aciertos al azar) es el que tradicionalmente ha presentado mayor imprecisión (Dinero y Haertel, 1977; Kolen, 1981; Lord, 1983; McKinley y Mills, 1985; Reckase, 1979; Ree, 1979; Thissen y Wainer, 1982; Van de Vijver, 1986; Wright, 1977a y b; Wright y Stone, 1979; Wainer y Wright, 1980; Yen, 1981). En un trabajo con datos simulados, Muñiz, Rogers y Swaminathan (1989), en concordancia con los datos anteriores, encontraron que los modelos resultaban muy robustos a violaciones relativas a los parámetros a, b y c. En concreto el Modelo de Rasch resultó altamente robusto a violaciones de $c=0$ y $a=k$, con lo que su uso tan habitual en la práctica, donde lo más probable es que dichos supuestos no se cumplan estrictamente, no constituiría un gran problema.

El análisis de los residuos, la comparación de las distribuciones estadísti-

cas y el uso de estadísticos como χ^2 son las técnicas habituales para el estudio del Ajuste de los modelos a los datos, echándose de menos una tecnología estadística más potente (Bock, 1972; Hambleton y Rogers, en prensa; Ludlow, 1985, 1986; Rogers y Hattie, 1987; Wright y Panchapakesan, 1969; Wright y Stone, 1979; Yen, 1981).

En suma, los modelos parecen razonablemente robustos, pero aún son necesarias investigaciones más concluyentes, especialmente acerca del parámetro c, por lo que los próximos años éste será uno de los campos de la TRI que más atención siga recibiendo por parte de los investigadores.

Función de Información

Sin duda alguna una de las aportaciones más relevantes, y que supone un cambio radical para la construcción y análisis de los instrumentos de medida, ha sido la Función de Información (FI), función que expresa la precisión con la que el test mide a lo largo de los distintos valores de la escala de la variable medida (Θ).

$$\text{Función de Información del test: } I(\Theta)=\Sigma\{[P'_i(\Theta)]^2/P_i(\Theta)Q_i(\Theta)\}$$

$$\text{Función de Información del ítem: } I(\Theta)=[P'_i(\Theta)]^2/P_i(\Theta)Q_i(\Theta)$$

En la TC se asumía que la cuantía del error de medida era la misma independientemente de la puntuación de los sujetos ($r_{ve}=0$), asunción que encajaba mal con numerosos datos indicadores de lo contrario (Lord, 1984). Piénsese, por ejemplo, lo poco razonable que resultaría asumir que un test muy difícil mida con la misma precisión a los sujetos altamente competentes y a los de muy baja competencia, obviamente discriminará mucho mejor entre los más competentes, llevado al extremo, entre los de baja competencia no discriminaría nada, todos sacarían un cero.

La FI del test (o del ítem) permitirá conocer para qué valores es más preciso el test (o el ítem), con lo que el uso de un test u otro va a depender del tipo de sujetos a evaluar; no se utilizará el mismo test con sujetos de competencia baja, media o alta, lo cual sólo es posible gracias a la invarianza de las mediciones respecto del test utilizado que permite la TRI. Nótese que a pesar de usar tests distintos las medidas de todos ellos estarán en la misma escala, incluso se puede conseguir que con la misma precisión. Surge así el concepto de test adaptado al sujeto, que aunque con sus orígenes en la TC, no podía operativizarse adecuadamente en aquel marco. La FI será, en suma, el gran avance en la modelización del error aportado por la TRI, pudiendo decirse que la época del mismo test para todos los sujetos ha terminado en la medición psicológica. Investigaciones actuales tratan de determinar con precisión en qué medida las variaciones muestrales extremas que capitalizan el error pueden afectar la configuración de la FI (Hambleton y Jones, 1991).

Bancos de Ítems

Los Bancos de Ítems (BI) constituyen la piedra angular de la TRI, sin ellos nada de lo dicho hasta ahora se mantendría y sería aplicable. No son una parte, ni una aplicación más de la TRI, se encuentran a la base de su formulación, lo cual no siempre se ha entendido del todo bien. *Los Bancos de Ítems contienen los ítems parametrizados (calibrados en la misma escala) que definen operativamente la variable medida.* El rasgo es el banco, los modelos no son responsables de las extrapolaciones, reificaciones y similares. La validez de contenido del banco determinará el grado de generalización de los resultados. Ésa es precisamente una de las razones por las que la TRI se ha impuesto más en áreas con dominios claramente definidos, como las educativas y profesionales, que en el ámbito de las variables psicológicas, menos operacionalizadas, aunque probablemente un esfuerzo en esta dirección evitaría muchas de las discusiones bizantinas en torno a las esencias.

En la actualidad existen numerosos BI de acceso internacional para varios campos profesionales y académicos, algunos de ellos con más de 100.000 ítems, que permiten al usuario adquirir por un módico precio el número de ítems que desee, pero, sobre todo, conformando esos ítems la *Función de Información Objetivo* que se desee a priori, según las necesidades de cada cual. En un mundo cada día más unificado y global los BI internacionales pueden jugar un importante papel en la extensión de los conocimientos científicos, que no saben de fronteras.

El programa MICROCAT para PC permite la construcción de BI, la aplicación por pantalla y la corrección y análisis de los resultados, tanto desde el punto de vista clásico (ITEMAN), como de la TRI (RASCAL, ASCAL).

Nótese que cuando se habla de invarianza de las mediciones respecto de los tests utilizados se está refiriendo a tests compuestos por ítems pertenecientes a un banco y calibrados en la misma escala, de lo contrario no hay tal. Puede decirse que sin la existencia de un banco de ítems la TRI no reporta ninguna ventaja fundamental frente a la TC.

Pueden consultarse buenas exposiciones sobre BI en *Applied Psychological Measurement* (1986), Gruijter y Van der Kamp (1976), Choppin (1976), Millman y Arter (1984), Wright y Bell (1984), Baker (1986), Masters y Evans (1986), Van Thiel y Zwarts (1989). Dado que el uso pertinente de los modelos de TRI conlleva la existencia de buenos bancos de ítems, la extensión y aprovechamiento de los modelos va a depender en gran medida del esfuerzo futuro en la construcción de estos bancos.

Tests a Medida

Los tests a medida fueron otro de los conceptos psicométricos potenciados por la onda expansiva de la TRI. Si bien sus antecedentes primarios pueden en-

contrarse ya en tests como el de Binet[5], en los que la secuencia que se presenta al niño depende de sus respuestas previas, estos tests encontrarán la solución idónea a sus problemas en el marco de la TRI. Los Test a Medida, también denominados tests adaptados al sujeto, tests de nivel flexible, tests ramificados, tests individualizados, tests programados, o tests secuenciales, tienen como característica definitoria fundamental que se construyen ajustados, adaptados, al nivel de los sujetos a los que se aplican, con lo que se gana en precisón (se minimiza el error de medida); consiguiéndose asimismo efectos colaterales deseables, como el aumento de la motivación de los sujetos, debido a que no se tienen que enfrentar a gran cantidad de ítems muy por encima de sus posibilidades, ni perder el tiempo contestando a ítems demasiado fáciles para su nivel. Además la longitud del test puede disminuirse manteniendo la misma fiabilidad. Estas mediciones ajustadas sólo eran posibles en un marco en el que las mediciones fuesen invariantes de los instrumentos, pues de poco valdría ajustar el test si luego las medidas no están en la misma métrica, y ese marco es lo que aportará la TRI. Por añadidura, los nuevos tiempos han traído el acceso masivo a los ordenadores, facilitando enormemente la implantanción práctica de este tipo de tests adaptados, lo que ha llevado a llamarlos a veces tests computerizados; si bien no conviene olvidar que el uso o no del ordenador no es una característica crucial, aunque deseable, de hecho existen cada vez más tests clásicos implementados en ordenador que no tienen nada que ver con los tests a medida, véase, por ejemplo, el número especial dedicado a los tests computerizados por la revista *Applied Psychology. An International Review*, o también, Baker (1989), Brzezinski y Hiscox (1984), Bunderson, Inouye y Olsen (1989).

Además del problema de la invarianza, la TRI vendrá a solucionar el otro problema central de los tests a medida, a saber, cómo elegir el test más adecuado para cada sujeto. El uso de la Función de Información en combinación con un banco de ítems permitirá elegir aquel test que maximice la información para los valores deseados de medida, es decir para los valores en los que se «estima» que se encuentra la competencia del sujeto a evaluar. Esa estimación previa puede hacerse siguiendo varias estrategias diferentes, destacando las así llamadas de *doble nivel y multinivel.* En la primera se aplica un test de *screening* previo que permite ubicar tentativamente la zona en la que se encuentra el sujeto, que posteriormente se evaluará con el test adaptado. En la estrategia multinivel, aunque caben otras alternativas, lo más habitual es ir aplicando ítem a ítem, progresando en dificultad o rebajando según las respuestas del sujeto. Este proceso plantea problemas técnicos interesantes, de por dónde empezar a evaluar, criterios de progreso-regreso, cuándo detener el proceso, etc. muy bien tratados por Lord (1980), también Kingsbury y Zara (1989). El programa MICROCAT (Assessment Systems Corporation, 1984) permite llevar a cabo todos los pasos descritos, incluido el análisis de los resultados con los modelos logísticos de 1, 2 y 3 parámetros, existiendo además

5. ...Quand l'on veut savoir sommairement si un enfant a l'intelligence de son âge, ou s'il est avancé ou en retard, il suffit de faire avec lui les épreuves de son âge; et l'exécution de ces épreuves ne prend certainement pas plus de 30 minutes qui seront coupées par 10 minutes repos, si on croit ce repos nécessaire a l'enfant... lorsqu'on a besoin d'une approximation plus grande, on fera beaucoup plus d'épreuves; si l'enfant a sept ans, on essayera les épreuves de huit, de neuf, de dix ans, par example (Binet y Simon, 1908, p. 82).

otro software disponible, una buena revisión del cual puede verse en Hsu y Yu (1989). La literatura actual sobre los tests a medida es abundante y floreciente. El texto editado por Wainer (1990) es ideal para una visión comprensiva, también la revisión de Hambleton, Zaal y Pieters (1991). Véase además, Urry (1977), Lord (1980), Weiss (1982, 1983, 1985), Green (1983, 1988), Green y cols., (1982, 1984), Wainer y Kiely (1987), Van der Linden y Zwarts (1989), Weiss y Kingsbury (1984), Reckase (1989).

Las ventajas citadas y la disponibilidad creciente de potentes microordenadores y terminales, el software adecuado y bancos de ítems, está potenciando cada vez más el uso de este tipo de tests, con notorias ventajas sobre los clásicos. Ello no quiere decir que todos los problemas estén resueltos, como es el caso de la correcta equiparación y de la seleección de ítems (Green, 1988), o problemas específicos de estimación de los parámetros (Wainer y Kiely, 1987). No obstante, no se corre mucho riesgo al predecir que los próximos años verán un gran crecimiento de los tests a medida y de las evaluaciones individualizadas a través de terminales y PC_S.

Tests Referidos al Criterio

El caso de los Tests Referidos al Criterio (TRC) es parecido al de los Tests a Medida que, aunque de uso conocido en psicometría, recibirán un nuevo impulso en el marco de la TRI, el cual permitirá resolver satisfactoriamente algunos de los principales problemas que tenían planteados. Como es bien sabido, se entiende por TRC aquellos tests cuyo interés genuino, al contrario que los normativos, se centra en la determinación del grado en el que los sujetos dominan ciertas materias o áreas de conocimiento, a las que se suele denominar criterio o dominio. Una buena panorámica del campo puede verse en Berk (1984), Hambleton (1982, 1990b), Hambleton y cols. (1978) o Popham (1978), además del número especial dedicado al tema por *Applied Psychological Measurement* en 1980.

La TRI viene a solucionar dos problemas básicos de los TRC, el primero, como no, la unificación de la métrica para los distintos tests utilizados (invarianza), y el segundo, la confección del test más adecuado para maximizar la información que aporta en torno al punto de corte entre los sujetos a clasificar como conocedores del área y los no conocedores. Es éste un problema más que de gran pedigree teórico, de suma importancia práctica, pues numerosas instituciones han de tomar decisiones de ese tipo en situaciones de lo más variado. Obviamente se trata de encontrar la prueba que discrimine al máximo en la zona donde se producirá la dicotomización, en términos de las TRI, que aporte la información máxima para el punto de corte. La problemática tecnológica implicada en la ubicación de estos puntos de corte ha generado toda una literatura especializada (Angoff, 1984; Berk, 1976, 1986; Chuang y cols., 1981; Livingston y Zieky, 1982; Shepard, 1980, 1984), así como el ajuste de la longitud del test a su eficiencia clasificadora (Wilcox, 1980; Hambleton, 1984; Crocker y Algina, 1986; Livingston y Zieky, 1982).

Dado su objetivo centrado en el dominio en vez de en las normas, los TRC se prestan más a ser utilizados en áreas donde éstos aparecen claramente definidos, como el campo educativo y profesional, más que el relativo a variables psicológicas como tales. La proliferación de bancos de ítems nacionales e internacionales agrupados por especialidades académico-profesionales ha supuesto un empuje para esta forma de medida, y es cada día más frecuente el ver expresadas las puntuaciones de los sujetos más que en relación a sus compañeros (tests normativos), referidas al porcentaje de materia que dominan. Un problema claro, por otra parte, es que ese tipo de variables de rendimiento es poco plausible considerarlas unidimensionales, por lo que la generalización de los modelos multidimensionales de TRI supondría una buena noticia para los TRC.

Equiparación de puntuaciones

La equiparación de las puntuaciones de los tests (*equating*) es la única solución dentro del marco clásico para establecer las equivalencias entre las puntuaciones de distintos tests midiendo la misma variable (con la misma precisión), una salida a la ausencia de invarianza. Sorprende, sin embargo, la poca ateción que se prestó a este aspecto en la literatura clásica, así Gulliksen (1950) lo trata de pasada, Lord y Novick (1968) apenas lo abordan, y los *Standards for Educational and Psychological Tests* de 1974 ni lo citan, teniendo que esperar hasta 1971 la llegada del clásico capítulo de Angoff (1971). A partir de esas fechas la literatura se multiplica (Holland y Rubin, 1982; Lord, 1980; *Standards for Educational and Psychological Testing*, 1985; Kolen, 1988; número especial de *Applied Psychological Measurement*, 1988).

Este desarrollo tardío tiene una explicación, mientras los tests más clásicos se aplicaban una y otra vez (el mismo test) a los sujetos, no había muchas exigencias de equiparación, los grupos normativos permitían la ubicación relativa en el grupo, y era más rentable pasar el mismo test que equiparar. Pero con el uso masivo de los tests, sobre todo para acceso a enseñanzas superiores, certificaciones, selección, etc., y un sinfín de academias entrenando para pruebas específicas, los tests se «queman» prácticamente tras una aplicación, máxime en países en los que es obligatorio publicar la prueba una vez calificada. Surge así la necesidad de la equiparación de los tests anteriores con los nuevos. Es como se ve un problema aplicado típico, más que de gran fuste téorico, si tal distinción tiene algún sentido. Para una revisión de los métodos clásicos utilizados, además del citado Angoff (1971), véase Skaggs y Lisitz (1986) y Petersen y cols. (1989). El problema de la equiparación queda automáticamente resuelto en el marco de la TRI, puesto que las mediciones son invariantes respecto del instrumento utilizado, no ha lugar equiparar, la métrica es común. Lo cual es cierto sensu stricto siempre que nos estemos refiriendo a un banco de ítems calibrados en la misma métrica y del cual se extraen los distintos tests utilizados, en ese caso los valores de θ estimados serán invariantes. Sin embargo, la puntuación empírica del sujeto

en el test dependerá siempre del test utilizado (que no la θ), y si por razones prácticas interesa hacer corresponder la puntuación empírica de un test (X_1) con su equivalente en otro (X_2), el instrumento ideal de equiparación automática es la Curva Característica del Test, que asignará para una misma θ estimada los correspondientes valores en cada test. Ahora bien, si el problema está resulto tan limpiamente en el marco de la TRI, ha de haber alguna razón por la que la equiparación de las puntuaciones siga siendo uno de los capítulos que más investigación sigue generando. La razón es sencilla, lo habitual por ahora no es disponer de inmensos bancos de ítems ya probados y calibrados de los que se van extrayendo tests a conveniencia, sino disponer de un banco de ítems calibrados y de un nuevo test que se desea expresar en la métrica del banco establecido. La equiparación alude al proceso de convertir el nuevo test a la métrica del banco. Nótese que ésta siempre será la situación cada vez que se desee añadir nuevos ítems al banco. Existen diversos procedimientos a tal efecto, siendo el más habitual la inclusión en el nuevo test de algunos ítems de anclaje ya calibrados en el banco, con el fin de establecer la correspondencia entre ambas métricas. Ése es el sentido de la equiparación en la TRI. Buenos tratamientos de la tecnología utilizada en el marco de la TRI pueden verse en Cook y Eignor (1983, 1989), Lord (1980), Hambleton y Swaminathan (1985), Lord y Wingersky (1983), Stocking y Lord (1983), Haebara (1980). De todo lo cual se colige de nuevo el papel central que juegan los bancos de ítems en la TRI.

Sesgo

Si bien de una forma u otra el concepto de sesgo en las pruebas psicológicas es tan antiguo como ellas, es cierto que tardará históricamente en recibir un tratamiento técnico. Clásicos como Gulliksen (1950), Lord y Novick (1968) o Thorndike (1971) prácticamente ni lo citan, y tampoco lo hace la edición de 1966 de los *Standards for Educational and Psychological Tests and Manuals.* Un factor estimulador de las investigaciones en torno al sesgo fueron sin duda los debates y polémicas de los años sesenta y setenta, y trabajos como el de Jensen (1969), quien precisamente publicará en 1980 un tratado enciclopédico sobre sesgo (Jensen, 1980), sometido a escrutinio detallado por los especialistas en la revista *Behavioral and Brain Sciences* (1980). En los últimos años abunda la literatura (Berk, 1982; Cole y Moss, 1989; Osterlind, 1983; Reynols y Brown, 1984; Shepard y cols. 1981, 1984, 1985), constituyendo probablemente el tópico en el que se concentra mayor número de investigaciones, a juzgar por las revistas del área, o por el último congreso de la AERA-NCME (en Chicago, 1991).

Se entiende en psicometría que un test (o ítem) está sesgado contra un grupo o clase de sujetos cuando este grupo o clase, a pesar de tener el mismo nivel que otros en la variable medida, sale sistemáticamente perjudicado al utilizar dicho test (o ítem). Nótese, que como es obvio, esta conceptualización no tiene nada que ver con la posibilidad de que ciertos grupos puntúen más bajo o más alto que otros en un test.

La definición adoptada sugiere inmediatamente la pregunta de cómo saber que dos sujetos o grupos tienen el mismo nivel en la variable medida, pues para saberlo hay que aplicar el test, y si está sesgado no hay manera de hacer inferencias cabales acerca del nivel de los sujetos. Esta circularidad no está bien resuelta en las técnicas clásicas para detectar el sesgo de los ítems, cuya lógica general consiste en representar los aciertos a cada ítem de los grupos a evaluar, frente a sus puntuaciones globales en el test y analizar las diferencias entre ámbos, mediante alguna técnica estadística, habitualmente χ^2 (Scheuneman, 1979; Ironson, 1982; Ironson y Subkoviack, 1979; Baker, 1981; Marasculo y Slaughter, 1981; Camilli, 1979). Una de las técnicas que actualmente se utiliza con mayor frecuencia es la propuesta por Holland y Thayer (1988), basada en el procedimiento de Mantel y Haenzel (1959), así como el Método Delta (Angoff y Ford, 1973; Angoff, 1982b). Sin olvidar que todos estos métodos estadísticos se llevan a cabo a posteriori, y que no eximen, mas aún, suponen, un escrutinio sistemático previo, muy bien descrito por Tittle (1982).

Los métodos citados funcionan bien en la práctica, con la sombra de fondo de que la estimación del nivel se hace con el test global, que implícitamente se asume insesgado, por lo que en puridad lo que realmente se comprueba con ellos es si un ítem está sesgado o no frente a los otros ítems tomados globalmente.

De nuevo la TRI vendrá con una solución novedosa, bajo la nueva óptica un ítem estará sesgado si su CCI es distinta para los grupos evaluados. Por tanto, se trataría de estimar las CCI para ambos (o más) grupos y ver en qué medida difieren. Se han propuesto varios métodos para estimar la significación estadística de estas discrepancias entre las CCI_S, desde la mera inspección visual hasta complejos análisis estadísticos (Berk, 1982; Hambleton y Swaminathan, 1985; Hambleton y Rogers, 1991; Lord, 1980; Rudner, 1977; Rosebaum, 1987; Rudner y cols., 1980; Shepard y cols., 1981, 1984, 1985; Hills, 1989; Mellenbergh, 1989; Raju, 1988).

La solución de la TRI parece perfecta salvado el problema técnico de encontrar un buen indicador de la significación estadística de la diferencia entre las CCI, pues aunque los métodos convergen razonablemente, cuando las diferencias entre las curvas son muy exiguas surgen algunas discrepancias entre unos y otros (Hoover y Kolen, 1984). A este problema técnico-estadístico habría que añadir una especie de contradicción teórica de fondo, a saber, si Θ es unidimensional y la estimación de los parámetros es invariante respecto de los sujetos como los modelos reclaman, resultaría teóricamente imposible que un mismo ítem generase dos CCI distintas, y si es así, una de dos, o Θ no es unidimensional, o la invarianza no se mantiene. Claro que la realidad es algo más boscosa que la teoría y raramente, si alguna vez, se hallarán unidimensionalidades perfectas, ni invarianzas milimétricas, probándose sistemáticamente con datos simulados que los métodos citados detectan con buena precisión los ítems sesgados. En suma, se dispone de una tecnología de eficacia práctica probada, pero las bases teóricas sobre las que se asienta no son todo lo sólidas que fuese de desear.

Conclusiones

Tras el somero repaso histórico y los breves comentarios sobre algunas de

las áreas en las que los modelos de Teoría de Respuesta a los Ítems han ejercido una mayor influencia, la conclusión más general es que estos modelos han supuesto el mayor avance de los últimos años en la medición psicológica y educativa. No son la panacea, ni todos sus problemas están totalmente resueltos, pero han introducido una nueva tecnología de la medida con claras aplicaciones en la práctica, algo de lo que otras modalidades propuestas históricamente han adolecido; de hecho una gran mayoría de los tests elaborados en la actualidad están construidos utilizando la TRI.

De cara al futuro cercano el uso de estos modelos se verá potenciado a medida que aumenten y se extiendan los bancos de ítems, imprescindibles para una aplicación coherente de los modelos. Otro segmento que reclama un mayor desarrollo es la elaboración de modelos unidimensionales para ítems de respuesta no sólo dicotómica, sino también politómica, a la vez que modelos multidimensionales para ambos tipos de respuesta, existiendo ya abundantes trabajos en esa dirección (Masters y Wright, 1984; Bock, 1972; Samejima, 1969, 1972, 1973, 1974; Embretson, 1984). Modelos de ese tipo permitirían cubrir amplias áreas de evaluación psico-educativa a las que no parece razonable exigir una unidimensionalidad perfecta y/o respuestas dicotómicas.

No se conocen con toda seguridad los efectos sobre el ajuste de los modelos de la violación del supuesto de unidimensionalidad, por lo que más investigación en este campo también será necesaria, dada la repercusión práctica que ello puede tener. Además, y traspasándolo todo, nunca se hará lo bastante para perfeccionar la estimación de los paramétros, y conocer con precisión las propiedades estadísticas de los estimadores, pues como señala McDonald (1989), a pesar de los grandes avances y aplicabilidad de la TRI, aspectos como la elección, estimación y ajuste de los modelos son todavía cuestiones abiertas.

Referencias

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

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Andrich, D. (1988). *Rasch models for measurement*. Beverly Hills, California: Sage.

Angoff, W.F. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Ansley, T.M. & Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two dimensional data. *Applied Psychological Measurement, 9*, 37-48.

Assessment Systems Corporation (1988). *User's manual for the Microcat testing system, version 3*. St. Paul, MN: Author.

Baker, F.B. (1965). Origins of the item parameters X_{50} and β as a modern item analysis technique. *Journal of Educational Measurment, 2*(2), 167-180.

Baker, F.B. (1977). Advances in item analysis. Review of Educational Research, 47(1), 151-178.

Baker, F.B. (1981). A criticism of Scheuneman's item bias technique. *Journal of Educational Measurement, 18*, 59-62.

Baker, F.B. (1985). *The basics of Item Response Theory*. Portsmouth, NH: Heinemann.

Baker, F.B. (1986). Item banking in computer based instructional systems. *Applied Psychological Measurement, 10*, 405-414.

Baker, F.B. (1987). Item parameter estimation under the one, two and three parameter logistic models. *Applied Psychological Measurement, 11*, 111-141.

Baker, F.B. (1989). Computer technology in test construction and processing. In R.L. Linn (Ed.), *Educational measurement* (pp. 409-428). New York: Macmillan.

Berger, M.P. & Knol, D.L. (1990). *On the assessment of dimensionality in multidimensional item responses theory models* (Research Report 90-8). Enschede: University of Twente, Department of Education.
Berk, R.A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 15*(4), 4-9.
Berk, R.A. (Ed.) (1982). *Handbook of methods for detecting test bias*. Baltimore, MD: The Johns Hopkins University Press.
Berk, R.A. (Ed.) (1984). *A guide to criterion-referenced test construction* (2ª Ed.). Baltimore, MD: The Johns Hopkins University Press.
Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*, 137-172.
Binet, A. & Simon, T.H. (1905a). Sur la nécessité d'établir un diagnostic scientifique des états inférieurs de l'intelligence. *L'Année Psychologique, 11*, 163-190.
Binet, A. & Simon, T.H. (1905b). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique, 11*, 191-244.
Binet, A. & Simon, T.H. (1908). Le développment de l'intelligence chez les enfants. *L'Année Psychologique, 14* 1-94.
Birbaum, A. (1957). *Efficient design and use of tests of ability for various decision-making problems* (Series Report nº 58-16, Project nº 7755-23). Randolph Air Force Base, TX: USAF School of Aviation Medicine.
Birnbaum, A. (1958a). *On the estimation of mental ability* (Series Report nº 15. Project nº 7755-23). Randolph Air Force Base, TX: USAF School of Aviation Medicine.
Birnbaum, A. (1958b). *Further considerations of efficiency in tests of a mental ability* (Technical Report nº 17. Project nº 7755-23). Randolph Air Force Base, Texas: USAF School of Aviation Medicine.
Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. En F.M. Lord y M. Novick, *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.
Bock, R.D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.
Bock, R.D. y Wood, R. (1971). Test theory. *Annual Review of Psychology, 22*, 193-224.
Brzezinski, E. & Hiscox, M. (Eds.) (1984). Microcomputers in educational measurement (Special issue). *Educational Measurement: Issues and Practice, 3*, 3-50.
Bunderson, C.V., Inouye, D.K. & Olsen, J.B. (1989). The four generations of computerized educational measurement. In R.L. Linn (Ed.): *Educational Measurement* (3rd ed.; pp. 467-408). New York: MacMillan.
Camilli, G. (1979). A critique of the chi-square method of assessing item bias (*Laboratory of Educational Research*). Boulder, CO: University of Colorado.
Carson, R.C. (1969). *Interaction concepts of personality*. Chicago: Aldine.
Choppin, B.H. (1976). Recent developments in item banking: a review. In D.N.M. Gruijter & L.J.T. Van der Kamp (Eds.), *Advances in psychological and educational measurement*. New York: Wiley.
Choppin, B.H. (1983). *A two parameter latent trait model* (CSE Report nº 197). Los Angeles, CA: UCLA, Center for the study of Evaluation.
Chuang, D.T., Chen, J.J. & Novick, M.R. (1981). Theory and practice for the use of cut scores for personal decisions. *Journal of Educational Statistics, 6*(2), 129-152.
Cole, N.S. & Moss, P.A. (1989). Bias in test use. In R.L. Linn (Ed.), *Educational Measurement* (pp. 201-219). New York: Mcmillan.
Cook, L.L. & Eignor, D.R. (1983). Practical considerations regarding the use of item response theory to equate tests. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 175-195). Vancouver, BC: Educational Research Institute of British Columbia.
Cook, L.L. & Eignor, D.R. (1989). Using item response theory in test score equating. *International Journal of Educational Research, 13*, 162-174.
Croker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston.
Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
Dinero, T.E. & Haertel, E. (1977). Applicability of the Rasch model with varying item discriminations. *Applied Psychological Measurement, 4*, 581-592.
Drasgow, F. (1982). Choice of test model for appropiateness measurement. *Applied Psychological Measurement, 6*, 297-308.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*, 77-90.

Drasgow, F., Levine, M.V. & McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-79.

Drasgow, F. & Parsons, C.K. (1983). Application of unidimensional item response theory models to multimensional data. *Applied Psychological Measurement, 7*, 189-199.

Ekman, G.E. & Sjoberg, L. (1965). Scaling. *Annual Review of Psychology, 16*, 451-474.

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

Feldt, L.S. & Brennan, R.L. (1989). Reliability. In R.L. Linn (Ed.), *Educational Measurement* (pp. 105-146). New York: Macmillan.

Ferguson, G.A. (1942). Item selection by the constant process. *Psychometrika, 7*, 19-29.

Folk, V.G. & Green, B.F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13*, 373-389.

Fraser, C. (1981). *NOHARM. A Fortran program for non-linear analysis by a robust method for estimating the parameters of 1, 2 and 3 parameter latent trait models*. Armidale, Australia: University of New England, Centre for Behavioral Studies in Education.

Goldstein, H. & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42*, 139-167.

Greaud, V.A. (1988, April). *Some effects of applying unidimensional IRT to multidimensional tests*. Paper presented at the AERA meeting, New Orleans.

Green, B.F. (1983). The promise of tailored tests. In H. Wainer y S. Messick (Eds.), *Principles of modern psychological measurement: A festschrift for Frederic M. Lord*. Hillsdale, NJ: LEA.

Green, B.F. (1988). Critical problems in computer based psychological measurement. *Applied Measurement in Education, 1*, 223-231.

Green, B.F., Bock, R.D., Humphreys, LG., Linn, R.B. & Reckase, M.D. (1982). *Evaluation plan for the computerized adaptive vocational aptitude battery*. Baltimore, MD: The Johns Hopkins University, Department of Psychology.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.B. & Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Gruijter, D.N.M. & van der Kamp, L.J.T. (Eds.) (1976). *Advances in psychological and educational measurement*. New York: Wiley.

Guilford, J.P. (1936). *Psychometric Methods*. New York: Wiley.

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

Haebara, T. (1980). Equating logistic ability scales by weighted least method. *Japanese Psychological Research, 22*, 144-149.

Hambleton, R.K. (1979). Latent trait models and their applications. In R. Traub (Ed.), *Methodological developments: new directions for testing and measurement* (pp. 13-32). San Francisco: Jossey-Bass.

Hambleton, R.K. (1982). Advances in criterion-referenced testing technology. In C.R. Reynolds & T.B. Gutkin (Eds.), *Handbook of school psychology*. New York: Wiley.

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

Hambleton, R.K. (1986). The changing conception of measurement: A commentary. *Applied Psychological Measurement, 10*, 415-421.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 147-200). New York: Macmillan.

Hambleton, R.K. (1990a). Item response theory: Introduction and bibliography. *Psicothema, 2*, 97-107.

Hambleton, R.K. (1990b). Criterion-referenced testing methods and practices. In T.B. Gutkin & C.R. Reynolds (Eds.), *The handbook of school psychology* (2ª ed.). New York: Wiley.

Hambleton, R.K. & Jones, R.W. (1991, April). *Influence of various factors on the accuracy of test information functions*. Paper presented at the meeting of the National Council on Measurement in Education, Chicago, Illinois.

Hambleton, R.K. & Rogers, H.J. (1991). Evaluation of the plot method for identifying potentially biased tests items. In P.L. Dann, S.H. Irvine, & J.M. Collis (Eds.), *Advances in computer-based human assessment*. Boston, MA: Kluwer Academic Publishers.

Hambleton, R.K. & Rogers, H.J. (en prensa). Promising directions for assessing item response model fit to test data. *Applied Psychological Measurement*.

Hambleton, R.K. & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*, 287-302.

Hambleton, R.K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Academic Publishers.
Hambleton, R.K., Swaminathan, H., Cook, L.L., Eignor, D.R. & Gifford, J.A. (1978). Developments in latent trait theory: models, technical issues, and applications. *Review of Educational Research, 48*, 467-510.
Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory*. Beverly Hills, CA: Sage.
Hambleton, R.K., Zaal, J.N. & Pieters, J.P.M. (1991). Computerized adaptive testing: Theory, applications, and standards. In R.K. Hambleton & J.N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and applications*. Boston, MA: Kluwer Academic Publischers.
Harrison, D.A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*(2), 91-115.
Hattie, J.A. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research, 19*, 49-78.
Hattie, J.A. (1985). Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.
Hills, J.R. (1989). Screening for potentially biased items in testing programs. *Educational Measurement: Issues and Practice, 8*, 5-10.
Holland, D.W. & Rubin, D.R. (1982). *Test equating*. New York: Academic Press.
Holland, P.D. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test validity*. Hillsdale, NJ: LEA.
Hoover, H.D. & Kolen, M.J. (1984). The reliability of six item bias indices. *Applied Psychological Measurement, 8*, 173-181.
Hsu, T.C. & Yu, L. (1989). Using computers to analyze item response data. *Educational Measurement: Issues and Practice, 8*, 21-27.
Hulin, C.L., Drasgow, F. & Parsons, C.K. (1983). *Item response theory. Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.
Ironson, G.H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore, MD: The Johns Hopkins University Press.
Ironson, G.H. & Subkoviak, M. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16*, 209-224.
Jensen, A.R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Research, 39*, 1-123.
Jensen, A.R. (1980). *Bias in mental testing*. New York: The Free Press.
Jones, L.V. & Appelbaum, M.I. (1989). Psychometric methods. *Annual Review of Psychology, 40*, 23-43.
Keats, J.A. (1967). Test theory. *Annual Review of Psychology, 18*, 217-238.
Kingsbury, G.G. & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measuremnt in Education, 2*, 359-375.
Kolen, M.J. (1989). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*, 1-11.
Kolen, M.J. (1989). Traditional equating methodology. *Educational Measurement: Issues and Practice, 7*, 29-36.
Krantz, D.H., Luce, R.D., Suppes, P. & Tversky, A. (1971). *Foundations of measurement*. New York: Academic Press.
Kuder, G.F. & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika 2*, 151-160.
Lawley, D.N. (1943). On problems connectes with item selection and test construction. *Proceedings of the Royal Society of Edinburg, 61*, 273-287.
Lawley, D.N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edinburg, 62*, 74-82.
Lazarsfeld, P.F. (1959). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer et al. (Eds.), *Measurement and prediction*. Princeton, NJ: Princeton University Press.
Levine, M.V. & Drasgow, F. (1982). Appropriateness measurement: Review, critique and validating studies. *British Journal of Mathematical and Statistical Psychology, 35*, 42-56.
Levine, M.V. & Rubin, D.B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.
Linn, R.L. (Ed.) (1989). *Educational measurement*. New York: MacMillan.
Linn, R.L. (1990). Has item response theory increased the validity of achievement test scores? *Applied Measurement in Education, 3*, 115-141.

Livingston, S.A. & Zieky, M.J. (1982). *Passing scores*. Princeton, NJ: Educational Testing Service.
Lord, F.M. (1952). A theory of test scores. *Psychometric Monographs*, n° 7.
Lord, F.M. (1953a). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13*, 517-549.
Lord, F.M. (1953b). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18*, 57-76.
Lord, F.M. (1977). Practical applications of item characteristic curve theory. *Journal of Ecucational Measurement, 14*, 117-138.
Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: LEA.
Lord, F.M. (1983). Small N justifies Rasch model. In R.J. Weiss (Ed.), *New horizons in testing*. New York: Academic Press.
Lord, F.M. (1984). Standard errors of measurement at different ability levels. *Journal of Educational Measurement, 21*, 239-243.
Lord, F.M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement, 23*, 157-162.
Lord, F.M. & Novick, M. (1968). *Statistical theories of mental tests scores*. Reading, MA: Addison-Wesley.
Lord, F.M. & Wingersky, M.S. (1983). *Comparison of IRT observed score and true-score equating* (Research Bulletin, 83-86). Princeton, NJ: Educational Testing Service.
Luce, R.D. & Narens, L. (1986). The mathematics underlying measurement on the continuum. *Science, 236*, 1527-1532.
Ludlow, L.H. (1985). A strategy for the graphical representation of Rasch model residuals. *Educational and Psychological Measurement, 45*, 851-859.
Ludlow, L.H. (1986). On the graphical analysis of item response theory residuals. *Applied Psychological Measurement, 10*, 217-229.
Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 715-748.
Marascuilo, L.A. & Slaughter, R.E. (1981). Statistical procedures for identifying possible sources of item bias based on chi-square statistics. *Journal of Educational Measurement, 18*, 229-248.
Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.
Masters, G.N. & Evans, J. (1986). Banking non-dichotomously scored items. *Applied Psychological Measurement, 10*, 355-367.
Masters, G.N. & Wright, B.D. (1984). The essential process in a family of measurement models. *Psychometrika, 49*, 529-544.
McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100-117.
McDonald, R.P. (1982). Linear versus non-linear models in item response theory. *Applied Psychological Measurement, 6*, 379-396.
McDonald, R.P. (1989). Future directions for item response theory. *International Journal of Educational Research, 13*, 205-220.
McKinley, R.L. & Mills, C.N. (1985). A comparison of several goodness of fit statistics. *Applied Psychological Measurement, 9*, 49-57.
Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 128-144.
Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13-104). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13-104). New York: MacMillan.
Michell, J. (1990). *An introduction to the logic of psychological measurement*. Hillsdale, NJ: LEA.
Millman, J. & Arter, J.A. (1984). Issues in item banking. *Journal of Educational Measurement, 21*, 315-330.
Mischel, W. (1968). *Personality and assessment*. New York: Wiley.
Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51*, 177-196.
Mislevy, R.J. & Bock, R.D. (1983). *BILOG: Item analysis and test scoring with binary logistic models* (Computer program). Mooresville, IN: Scientific Software, Inc.
Mislevy, R.J. & Stocking, M.L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement, 13*, 57-75.
Mosier, C.I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. *Psychological Review, 47*, 355-366.
Mosier, C.I. (1941). Psychophysics and mental test theory II. The constant process. *Psychological Review, 48*, 235-249.
Muñiz, J. (1990). *Teoría de respuesta a los ítems*. Madrid: Pirámide.
Muñiz, J. (1991). *Introducción a los métodos psicofísicos*. Barcelona: PPU.
Muñiz, J., Rogers, H.J. y Swaminathan, H. (1989). Robustez de las estimaciones del modelo de Rasch

en presencia de aciertos al azar y discriminación variable de los ítems. *Anuario de Psicología, 43*, 83-97.
Narens, L. (1985). *Abstract measurement theory*. Cambridge, MA: MIT Press.
Narens, L. & Luce, R.D. (1986). Measurement: The theory of numerical assignment. *Psychological Bulletin, 99*, 166-180.
Osterlind, S.J. (1983). *Test item bias*. Beverly Hills, CA: Sage.
Petersen, N.S., Kolen, M.J. & Hoover, H.D. (1989). Scaling, norming, and equating. In R.L. Linn (Ed.), *Educational measurement* (pp. 221-262). New York: Macmillan.
Pfanzagl, J. (1968). *Theory of measurement*. New York: Wiley.
Popham, W.J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice Hall.
Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute for Educational Research.
Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*, 207-230.
Reckase, M.D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8*, 11-15.
Reckase, M.D., Ackerman, T.A. & Carlson, J.E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement, 25*, 193-203.
Ree, J.M. (1979). Estimating item characteristic curves. *Applied Psychological Measurement, 3*, 371-385.
Reynolds, C.R. & Brown, R.T. (Eds.) (1984). *Perspectives on bias in mental testing*. New York: Plenum Press.
Richardson, M.W. (1936). The relation between the difficulty and the differential validity of a test. *Psychometrika, 1*, 33-49.
Roberts, F. (1979). *Measurement Theory*. Reading, MA: Addison-Wesley.
Rogers, H.J. & Hattie, J.A. (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*, 47-58.
Rosenbaum, P.R. (1987). Comparing item characteristic curves. *Psychometrika, 52*, 217-233.
Rudner, L.M. (1977, April). *An approach to biased item identification using latent trait measurement theory*. Paper presented at the meeting of AERA, New York.
Rudner, L.M., Getson, P.R. & Knight, ,D.L. (1980). A Monte Carlo comparison of seven biased item detection techniques. *Journal of Educational Measurement, 17*, 1-10.
Samejima, F. (1969). Estimation of latent ability using a response pattern of grade scores. *Psychometric Monographs*, nº 17.
Samejima, F. (1972). A general model for free response data. *Psychometric Monographs*, nº 18.
Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38*, 203-219.
Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. *Psychometrika, 39*, 11-121.
Savage, L.W. & Ehrlich, P. (Eds.) (1990). *Philosophical and foundational issues in measurement theory*. Hillsdale, NJ: LEA.
Scheuneman, J. (1979). A method of assessing bias in test items. *Journal of Educational Measurement, 16*, 143-152.
Shepard, L.A. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4*, 447-467.
Shepard, L.A. (1984). Setting performance standards. In R. Berk (Ed.), *A guide to criterion-referenced test construction*. Baltimore, MD: The Johns Hopkins University Press.
Shepard, L.A., Camilli, G. & Averill, M. (1981). Comparison of procedures for detecting test-item bias with both internal and external ability criteria. *Journal of Educational Statistics, 6*, 317-375.
Shepard, L.A., Camilli, G. & Williams, D.M. (1984). *Journal of Educational Statistics, 9*, 93-128.
Shepard, L.A., Camilli, G. & Williams, D.M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77-105.
Skaggs, G. & Lissitz, R.W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*, 495-529.
Spearmen, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.
Standards for educational and psychological testing. Washington, DC: American Psychological Association.
Stanley, J.C. (1971). Reliability. In R.L. Thorndike (Ed.), *Educational measurement*. Washington, DC: American Council on Education.
Stocking, M. & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.
Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589-617.

Swaminathan, H. (1983). Parameter estimation in item response models. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 24-44). Vancouver, BC: Educational Research Institute of British Columbia.
Swaminathan, H. & Gifford, J.A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics, 7*, 175-192.
Swaminathan, H. & Gifford, J.A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika, 50*, 349-364.
Swaminathan, H. & Gifford, J.A. (1986). Bayesian estimation in the three-parameter logistic model. *Psychometrika, 51*, 589-601.
Sympson, J.B. (1983). *A new IRT model for calibrating multiple choice items*. Paper presented at the meeting of the Psychometric Society, Los Angeles, California (citado de Thissen y Steinberg, 1986).
Thissen, D.M. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 50*, 349-364.
Thissen, D.M. (1983). *MULTILOG: Item analysis and scoring with multiple category response models*. Chicago, IL: International Educational Services.
Thissen, D.M. & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49*, 501-519.
Thissen, D.M. & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*, 567-577.
Thissen, D.M. & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104*, 385-395.
Thissen, D.M. & Wainer, H. (1982). Some standards errors in item response theory. *Psychometrika, 47*, 397-412.
Thorndike, R.L. (Ed.) (1971). *Educational measurement*. Washington, DC: American Council on Education.
Thurstone, L.L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology, 16*, 433-451.
Thurstone, L.L. (1927). The unit of measurement in educational scales. *The Journal of Educational Psychology, 18*, 505-524.
Thurstone, L.L. (1928a). The absolute zero in intelligence measurement. *The Psychological Review, 35*, 175-197.
Thurstone, L.L. (1982b). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.
Thurstone, L.L. & Ackerson, L. (1929). The mental growth curve for the Binet tests. *The Journal of Educational Psychology, 20*, 569-583.
Tittle, C.K. (1982). Use of judgmental methods in item bias studies. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore, MD: The Johns Hopkins University Press.
Torgerson, W,S, (1958), *Theory and methods of scaling*. New York: Wiley.
Traub, R.E. & Lam, Y.R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology, 36*, 19-48.
Tsutakana, R.K. & Lin, H. (1986). Bayesian estimation of item response curves. *Psychometrika, 51*, 251-268.
Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1-13.
Tucker, L.R. (1963). Scaling and test theory. *Annual Review of Pscychology, 14*, 351-364.
Tucker, L.R. (1987). *Developments in classical item analysis methods* (ETS Research Report 87-46). Princeton, NJ: Educational Testing Service.
Urry, V.W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.
Vale, J.R. & Vale, C.A. (1969). Individual differences and general laws in psychology: A reconciliation. *American Psychologist, 24*, 1093-1108.
Van de Vijver, F.J. (1986). The robustness of Rasch estimates. *Applied Psychological Measurement, 10*, 45-57.
Van der Linden, W.J. (1986). The changing conception of measurement in education and psychology. *Applied Psychological Measurement, 10*, 325-332.
Van der Linden, W.J. (1989). Some procedures for computerized ability testing. *International Journal of Educational Research, 13*, 176-188.
Van Thiel, C.C. & Zwarts, M.A. (1986). Development of a testing service system. *Applied Psychological Measurement, 10*, 391-403.
Wainer, H. (Ed.) (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: LEA.
Wainer, H. & Braun, H.I. (Eds.) (1988). *Test validity*. Hillsdale, NJ: LEA.
Wainer, H. & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.
Wainer, H. & Wright, B.D. (1980). Robust estimation of ability in the Rasch model. *Psychometrika, 45*, 373-391.
Way, W.D., Ansley, T.N. & Forsyth, R.A. (1988). The comparative effects of compensatory and noncom-

pensatory two-dimensional data in unidimensional IRT estimates. *Applied Psychological Measurement, 12*, 239-252.
Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.
Weiss, D.J. (Ed.) (1983). *New horizons in testing*. New York: Academic Press.
Weiss, D.J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*, 774-789.
Weiss, D.J. & Davison, M.L. (1981). Test theory and methods. *Annual Review of Psychology, 32*, 629-658.
Weiss, D.J. & Kingsbury, G.G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*, 361-375.
Whitley, S. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.
Wilcox, R.R. (1980). Determining the length of a criterion-referenced test. *Applied Psychological Measurement, 4*, 425-446.
Wingersky, M.S., Barton, M.A. & Lord, F.M. (1982). *LOGIST 5.0, version 1.0 user's guide*. Princeton, NJ: Educational Testing Service.
Wingersky, M.S., Thissen, D. & Wainer, H. (1983). *Estimation of the form of the item characteristic curve using monotone splines*. Paper presented at the meeting of the Psychometric Society, Los Angeles, California (citado de Thissen y Steinberg, 1986).
Wood, R., Wingersky, MS. & Lord, F.M. (1976). *LOGIST: a computer program for estimating examinee ability and item characteristic curve parameters* (Research Report 76-6). Princeton, NJ: Educational Testing Srevice.
Wright, B.D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, NJ: Educational Testing Service.
Wright, B.D. (1977a). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.
Wright, B.D. (1977b). Misunderstanding of the Rasch model. *Journal of Educational Measurement, 14*, 219-226.
Wright, B.D. & Bell, S.R. (1984). Item banks: What, why, how. *Journal of Educational Measurement, 21*, 331-346.
Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.
Wright, B.D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.
Wright, B.D. & Stone, M.H. (1979). *Best test design*. Chicago. MESA Press.
Wright, B.D. & Mead, R.J. (1976). *BICAL. Calibrating rating scales with the Rasch model* (Research memorandum n° 23). Chicago: Statistical Laboratory, Department of Education, University of Chicago.
Wright, B.D., Mead, R.J. & Bell, S.R. (1979). *BICAL: A Rasch program for the analysis of dichotomous data*. Chicago: MESA Press.
Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5*, 245-262.
Yen, W.M. (1984). Effects of item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125-145.
Yen, W.M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291.
Zwick, R. (1987). Assessing the dimensionality of NAEP reading data. *Journal of Educational Measurement, 24*. 293-308.