Psicothema
2000. Vol. 12, Supl. nº 2, pp. 25-29

ISSN 0214 - 9915 CODEN PSOTEG

Perspectivas cognitivas y modelos TRI en el desarrollo de tests de rendimiento: una exploración preliminar de funciones de información

Salvador Algarabel, Eva Rosa, Juan Carlos Ruiz, Carmen Dasí y Alfonso Pitarque
Universitat de València

Los tests de rendimiento son punteros en incorporar innovaciones psicométricas, necesarias para medir la concepción actual de rendimiento. Dicho concepto ha ido cambiando desde el conductismo, que defendía la medición de muestras de conducta observables, sin apenas referencias a procesos internos, hasta ahora, en que desde la Psicología Cognitiva y las corrientes educativas, se defiende la necesidad de explorar también los mecanismos internos que se ponen en funcionamiento ante una prueba de rendimiento. Se trata de intentar aproximarse a la estructura mental que el examinado tiene de un dominio de conocimiento determinado. La teoría de respuesta al ítem se ha convertido en el instrumento básico actual para escalar el conocimiento individual. Presentamos un estudio preliminar en el que calculamos las funciones de información de varios modelos de escalamiento, aplicados a la evaluación del conocimiento complejo. Nuestra aproximación utiliza el concepto de «mapa conceptual», usado en la evaluación propuesta por las nuevas corrientes educativas, pero intentando sustituir sus aspectos cualitativos por un escalamiento estandarizado parecido al de los tests tradicionales. Para ello utilizamos una escala graduada de respuesta, y analizamos las funciones de información de varios modelos politómicos, y uno dicotómico (modelo de dos parámetros), como línea base de comparación. Los resultados indican que el modelo de crédito parcial generalizado es adecuado para escalar el rendimiento entendido de esta forma.

Cognitive perspectives and IRT models in the development of achievement tests: A preliminary exploration of information functions. Achievement tests are good tools to implement the new psychometric technologies which are playing an important role in the modern definition of achievement. The view of achievement has changed from the behaviourism, which used to view it as the measurement of overt behaviours without reference to internal cognitive processes, until nowadays in which from both, Cognitive Psychology and Performance Assessment movements, the need to explore the internal processes brought into play by the examinee is considered as necessary. From this current view achievement is assimilated to the measurement of the mental structure of a knowledge structure. Item response theory is the basic instrument for scaling individual knowledge. This is a preliminary study in which we explore the information functions of several IRT models as scaling models of complex knowledge. We standardize the concept of mental model, coming from performance assessment, asking a relatedness (5 point scale) judgement from the examinee for all pairs of a series of concepts. In this work we explore and compare the information functions of several politomous models, and one dichotomous model (two parameter model). Results show that generalised partial credit model is appropriate to measure this new conception of achievement.

Los tests de rendimiento actuales son con diferencia los instrumentos psicométricos más dinámicos debido a que tienen que renovarse continuamente, y, por lo tanto, existen mayores oportunidades de incorporar los avances metodológicos. Esta afirmación es cierta para los Estados Unidos, ya que debido a los pases masivos que el Educational Testing Service tiene que realizar de tests de rendimiento y aptitudes a prácticamente toda la población estudiantil, la investigación en este campo es, con diferencia, mucho mayor que en otros más enfocados al diagnóstico y tratamiento de patologías psicológicas o incluso de medición pura de las aptitudes.

El desarrollo de un test de rendimiento supone la consideración cuidadosa de, al menos, tres aspectos esenciales. El primero y el segundo tienen que ver con la representación de constructo y con la amplitud nomotética (por ej. Embretson, 1983; Prieto y Delgado, 1999), y el tercero, con la elección de un modelo para la representación de la información (por ej., Muñiz, 1996; van der Linden y Hambleton, 1997). Los dos primeros aspectos desembocan en la estructura de validez del test y están asociados con la investigación psicológica y educativa, mientras que el tercero tiene que ver de lleno con los instrumentos psicométricos de representación del conocimiento disponibles. En las dos últimas décadas ha habi-

Correspondencia: Salvador Algarabel
Facultad de Psicología
Universitat de València
46010 Valencia (Spain)
E-mail: algarabe@uv.es

do cambios notables tanto en la definición psicológica del «rendimiento» como en el desarrollo de modelos de representación, en una medida que merece una reconsideración por aquellos que llevamos a cabo investigación en la representación y medida del conocimiento. Tres maneras diferentes de definir un dominio de conocimiento van a ser consideradas: el enfoque conductual, el enfoque cognitivo y, finalmente, el educativo, representado por la llamada evaluación basada en conducta.

La definición conductual de la adquisición de conocimiento, prevalente en psicología teórica hasta mediados de los 70, pero aún duradera en algunos campos aplicados educativos, acostumbra a definir la adquisición en términos de fenómenos observables, no haciendo apenas referencia a procesos internos. Cuando se postula algún proceso encubierto se habla de «hábitos», «tendencias de respuesta», y en general, variables hipotéticas de tipo aditivo. Una forma clara de esta forma de concebir la adquisición se refleja (o reflejó, debido a su bajo uso actual) en la llamada «enseñanza programada», en la que el conocimiento se fragmenta en pequeñas unidades que se van adquiriendo de forma incremental (por ej., Holland y Skinner, 1973). En realidad, en este planteamiento un dominio de conocimiento se segmenta en una serie de estímulos que producen una respuesta en el examinado. Así pues, el enfoque conductual consiste en definir y fragmentar el universo de contenidos, para posteriormente proceder a llevar a cabo un muestreo representativo del mismo y una enseñanza o adquisición de la secuencia de elementos de información en el orden adecuado en función de su dificultad.

En contraste, la aproximación cognitiva a la adquisición considera el dominio de rendimiento en función de unas habilidades cognitivas que hacen referencia a unos procesos internos asociados fundamentalmente con la estructura de la memoria (esquemas, modelos mentales), y los procesos de inferencia. Este enfoque teórico es dominante en la psicología experimental y teórica actual y tiene un gran peso en la psicología y movimientos educativos. Desde este punto de vista, el rendimiento debe ser entendido como un constructo que debe referirse a diferentes niveles de la adquisición del conocimiento. El resultado final es el conocimiento tal como lo tiene un experto, a saber, un conjunto estructurado de modelos mentales construido a través de largas sesiones de práctica, que le permite poner en funcionamiento sofisticadas estrategias y grandes volúmenes de conocimiento de forma automatizada. Mediante la práctica se consigue sobrepasar las limitaciones impuestas por la memoria de trabajo (Ericsson, 1996).

En este contexto, el rendimiento puede ser definido (Niemi, 1999) como el dominio de «los conceptos y principios fundamentales, hechos y proposiciones importantes, esquemas, conocimiento estratégico e integración del conocimiento».

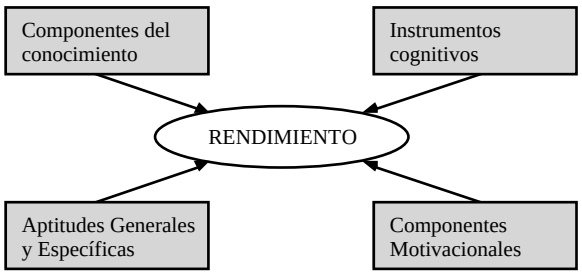

Figura 1. *Principales aspectos determinantes del rendimiento según Ruiz-Primo (1998)*

El siguiente esquema es un reflejo de las teorías cognitivas en el campo educativo directamente orientado a discutir los problemas de la evaluación del rendimiento (tomado con modificaciones de Ruiz-Primo, 1998).

El rendimiento se ve determinado, a efectos del presente trabajo, por los componentes de conocimiento, por instrumentos cognitivos y por las aptitudes generales y específicas. Los «instrumentos cognitivos» son simplemente la capacidad del sistema por monitorizar y planificar el proceso de aplicación de solución de problemas. Los componentes del conocimiento son: conocimiento declarativo (contenidos específicos del dominio), conocimiento procedimental (sistemas de producción específicos al dominio) y, finalmente, conocimiento estratégico (esquemas y estrategias de solución de problemas).

En el movimiento educativo de «evaluación del rendimiento» (performance assessment) se asume, tal como se presenta en las tablas (Ruiz-Primo, 1998), que los procesos cognitivos anteriores se evalúan en función de diversos procedimientos. Particularmente interesante, es la constatación de que entre los teóricos del campo educativo se considera de forma generalizada que el conocimiento complejo sólo puede evaluarse por medio de respuestas elaboradas, lo que antiguamente se hacía por medio de métodos tradicionales.

En resumen, el rendimiento es la competencia de una persona en relación con un dominio de conocimiento, que es lo que se recoge en los nuevos Estándares para la construcción de tests («Standards for Educational and Psychological Tests», 1999), en los que literalmente se dice que las mediciones del rendimiento son «descripciones de la competencia de un adulto o un estudiante en un área de contenido particular, normalmente definida como un conjunto de categorías ordenadas a lo largo de un continuo».

Formato de los items y evaluación del rendimiento

El enfoque «conductual» siempre ha sido proponente del ítem de elección múltiple. Históricamente, muchos dominios de conocimiento han sido analizados desde el punto de vista conductual, incluyendo áreas de conocimiento psicológicas como «el análisis de la conducta». Sin embargo, el ítem de elección múltiple también ha sido considerado como el elemento básico de sondeo de la información en el movimiento cognitivo. En este caso puede hacerse complejo, como en aquellos items de elección múltiple con calculadora, e incluso automatizar la corrección, como en algunos de los nuevos ítems utilizados en los tests que administra el Educational Testing Service.

Psicométricamente, el ítem de elección múltiple tiene numerosas ventajas, siendo la principal de ellas la estandarización. Esto es, una enorme cantidad de variables indeseables que pueden influir

Tabla 1
Tipos de ítems que pueden sondear los distintos tipos de conocimiento según Ruiz-Primo (1998)

	Conocimiento Declarativo	Conocimiento Procedimental	Conocimiento Estratégico
Amplitud	- Elección Múltiple - Respuesta corta	- «Performance Assessment»	- «Performance Assessment»
Estructura	- Mapas Conceptuales	- Mapas Conceptuales	- Mapas Conceptuales

sobre la respuesta son eliminadas suministrando las diversas alternativas. El procedimiento es además económico en tiempo y coste. En un estudio de Wainer y Thissen (1993) se demuestra que para conseguir un determinado nivel de fiabilidad el coste y el tiempo de prueba es inferior en el caso del ítem de elección múltiple.

Sin embargo, entre los inconvenientes están las críticas al modelo psicológico que supuestamente sustenta. Gran parte de los teóricos educativos considera al ítem de elección múltiple como «sondeador» de amplitud de información más que de conocimiento estructural o de otro tipo de procesos cognitivos complejos. Adicionalmente, el ítem de elección múltiple tiene una precisión menor porque aún cuando requiera procesos cognitivos complejos, éstos quedan opacos en la solución correcta. Esto es, el rendimiento es algo más que soluciones correctas o incorrectas, añaden los proponentes de la evaluación basada en el rendimiento.

Los pedagogos orientados a la evaluación «auténtica» consideran que «los procesos» que deben medirse para mostrar la adquisición de conocimiento no se pueden muestrear con un test estandarizado de corte clásico porque el ítem de elección no permite reflejar estos procesos. De hecho, gran parte de los movimientos educativos contemporáneos basan la evaluación en las preguntas tradicionales de respuesta abierta, en las que existe un evaluador o evaluadores que la enjuician y llegan a una puntuación final. Esta posición teórica actúa como motor para la generación de procedimientos automáticos de corrección que ya están implementándose prácticamente en el Educational Testing Service (por ej. Foltz, 1996; Foltz, Kintsch y Landauer, 1998; Landauer, Foltz, y Laham, 1998).

Por otro lado, hay que tener en cuenta que la Psicología Básica de corte experimental lleva largo tiempo investigando los procesos de representación del conocimiento, aunque no orientada a la escalación de personas. Puesto que es un tipo de psicología orientada al estímulo, los procedimientos de análisis utilizados han sido el escalamiento multidimensional, la aglomeración jerárquica, árboles aditivos o pathfinder. Puede parecer paradójico que entre los procedimientos de investigación empírica no figure ninguno asociado con modelos de medida utilizados para el desarrollo de tests, pero esa es la realidad, y este es un punto que merece una atención especial por los investigadores que trabajan en este campo.

Modelos de medida para evaluar rendimiento

El concepto de rendimiento, tal como ha sido definido, es un constructo heterogéneo, que aunque puede ser fácilmente dividido en componentes, las habilidades que un individuo demuestra por separado en cada uno de estos componentes suelen estar muy interrelacionadas. Todo esto provoca cierta dificultad en encontrar modelos de medida que sean capaces de captar en su totalidad la amplitud del concepto.

En la actualidad, los modelos psicométricos de uso más generalizado son los basados en la Teoría de Respuesta al Item, que son los únicos que van a ser considerados en este trabajo. Estos modelos, que surgen con fuerza a partir de los 80 (p.e. Lord, 1980) para salvar los problemas, sobre todo de invariancia de estimaciones, que presentaba la Teoría Clásica de los Tests, asumen que la probabilidad de responder correctamente a un ítem, dado un nivel de habilidad del sujeto, es una función no lineal (logística o normal) de la dificultad, la discriminación y la adivinación del ítem, dependiendo del modelo especificado. Una revisión de los modelos TRI excedería con mucho el objetivo de este trabajo dada la gran proliferación de los mismos en los últimos años (ver p.e. van der Linden y Hambleton, 1997). Los modelos TRI constituyen una clase de modelos de medida especialmente convenientes para escalar tanto los procesos inferenciales como los contenidos del conocimiento, aunque en el test típico ambos aspectos del rendimiento son indistinguibles.

Los datos preliminares que aquí presentamos son parte del análisis exploratorio inicial de modelos de TRI que mejor ajustan datos cuya significación psicológica está asociada con la adquisición de conocimiento complejo, sujeto a diversas condiciones experimentales. En este trabajo se presentan datos que indican que los modelos TRI, particularmente los modelos Rasch, en general, tal como el modelo de crédito parcial, o no Rasch, tal como el modelo de crédito parcial generalizado (Muraki, 1992) son instrumentos eficacísimos en el escalamiento del conocimiento desde una perspectiva cognitiva.

Método

Se realizó un experimento con el objetivo de comparar la eficacia de varios modelos TRI (funciones de información) en la valoración del rendimiento de los sujetos experimentales en un dominio específico de conocimiento en el que habían sido instruidos. Los modelos comparados fueron: modelo de dos parámetros, modelo nominal, modelo graduado y modelo de crédito parcial generalizado.

Se manipuló el tipo de instrucción que habían recibido tres grupos de alumnos en la asignatura de «Diseños de Investigación Experimental». Aunque todos ellos recibían las mismas explicaciones teóricas de los conceptos fundamentales de la materia, en las clases de práctica, los alumnos realizaban ejercicios sobre problemas, bien con contenido psicológico, o bien de cálculo estadístico, o bien una mezcla de ambos tipos. El conocimiento complejo en este caso se midió de acuerdo con el grado de relacionalidad (escala de 5 puntos) percibida entre pares de conceptos relacionados con un artículo de investigación leído por los alumnos. Estos conceptos eran de naturaleza puramente estadística (por ej. hipótesis nula) o de naturaleza puramente psicológica (por ej. tipo de información, que era una variable independiente en el artículo leído), emparejándose todos ellos entre sí.

Se midió el rendimiento a través del grado de acuerdo entre los juicios emitidos por los alumnos y los juicios del experto, que en este caso era el profesor. Este tipo de medición del conocimiento es una alternativa estandarizada a los mapas conceptuales cualitativos utilizados por los teóricos educativos asociados con el movimiento de evaluación del rendimiento. Los modelos, cuyas funciones de información se están comparando, son el modelo de dos parámetros (para analizar la posible ganancia en información asociada con los modelos politómicos), el modelo nominal, el modelo politómico graduado de Samejima, y el modelo de crédito parcial generalizado (Muraki, 1992). La descripción matemática de estos modelos no se va a hacer aquí por espacio, pero puede encontrarse en cualquier manual psicométrico (por ej. Van der Linden y Hambleton, 1997).

Resultados

Los gráficos siguientes presentan las funciones de información calculadas sobre los datos obtenidos para los cuatro modelos comparados.

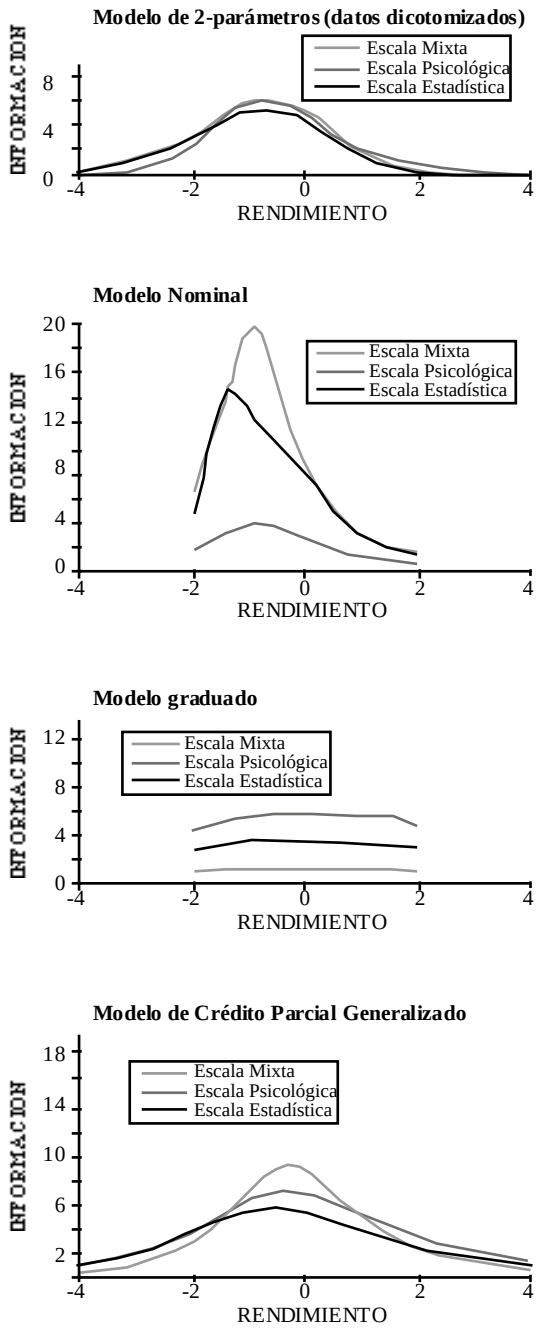

Figura 2. *Funciones de información obtenidas para cuatro modelos TRI*

Conclusiones

Podría pensarse que la relación entre la estructura psicológica del modelo que se está implementando y cualquier modelo concreto es unívoca. Esta relación no es de esta forma por lo que la elección de modelo se debe hacer en función de una consideración diversa de criterios. Puede apreciarse que si la respuesta lo permite, un modelo politómico ofrece más información que uno dicotómico, indicando un aumento de precisión.

Las funciones de información para el modelo nominal son mayores que para cualquier otro. Hay que tener en cuenta, sin embargo, que este modelo no es ordenado, y que el hecho de que se comparen categorías que están muy separadas entre sí (categorías no contiguas), puede hacer que la precisión sea mayor. Sin embargo, esta precisión no está asociada con la psicología del modelo representado, puesto que el ordenamiento de las categorías es un aspecto importante de la respuesta. Por lo tanto, parece que la mejor elección se encuentra entre el modelo graduado de Samejima y el modelo de crédito parcial generalizado (Muraki, 1992). Obsérvense las funciones de información de ambos modelos. En el caso de Samejima el escalamiento produce menor información. Por ello, parece recomendable el modelo de crédito parcial generalizado ya que, en función de los análisis previos, parece que los ítems difieren más en su poder discriminativo. En este modelo, que preferimos expresar a partir del modelo lineal generalizado de Mellenbergh (1994) con categorías adyacentes, la plausibilidad de elegir una categoría sobre otra viene dada por, siendo «b» el parámetro de dificultad, «a» el de discriminación y θ la habilidad:

$$\ln(\tau_{ijk+1} / \tau_{ijk}) = b_{ik+1} - b_{ik} + (a_{ik+1} - a_{ik})\theta_j = b'_{ik} + a'_{ik}\theta_j$$

En este caso,

$$b'_{ik} = b_{ik+1} - b_{ik} \quad \text{y} \quad a'_{ik} = a_{ik+1} - a_{ik}$$

El parámetro b'_{ik} es específico a un ítem y a una categoría, y se descompone en parámetros separados para ítem y para categoría. Esto es,

$$b'_{ik} = d_i + e_k$$

en donde la suma de los parámetros de las categorías es cero (los e's) y las pendientes (a's) intraítem son iguales. La estimación de los parámetros de los ítems se realizó a partir del modelo de crédito parcial generalizado, en base al método de máxima verosimilitud marginal. El análisis detallado de los datos del estudio nos permitirá comparar el escalamiento producido por el modelo de crédito parcial generalizado con el producido por otros procedimientos extendidos en el campo de la psicología experimental humana, no orientados al escalamiento de personas, tal como el escalamiento multidimensional. Esta comparación nos permitirá saber si, al margen de los problemas asociados con el tamaño de las muestras, los modelos TRI pueden ser buenos como instrumentos generales para representar el conocimiento.

Agradecimientos

Esta investigación fue financiada por una beca de la Dirección General de Investigación Científica y Técnica (PB/97-1379) del Ministerio de Educación y Ciencia.

Referencias

American Psychological Association, American Educational Research Association y Nacional Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing.* Washington, DC.

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

Ericsson, K. A. (1996).The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson, ed., *The road to excellence. The acquisition of expert performance in the arts and sciences, sports and games*, Mahwah, N. J.: Erlbaum, 1-50.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28*, 197-202.

Foltz, P. W., Kintsch, W., y Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes, 25*, 285-307.

Holland, J. G., y Skinner, B. F. (1973). *Análisis de conducta*. México: Trillas (original, 1961).

Landauer, T. R., Foltz, P. W., y Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

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

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115*, 300-307.

Muñiz, J. (1996). *Psicometría*. Madrid: Universitas.

Muraki, E. (1992). A generalizad partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Niemi, D. (1999, February). *Assessment models for aligning standards and classroom practice*. UCLA Graduate School of Education and Information Studies. Center for the Study of Evaluation. National Center for Research on Evaluation, Standards and Student Testing. Conference of The American Association of School Administrators.

Prieto, G. y Delgado, A.R. (1999): Medición cognitiva de las aptitudes. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests informatizados: Fundamentos y aplicaciones*. Madrid: Pirámide.

Ruiz-Primo, M. A. (1998). *Science ahievement: What we have learned from two alternative assessments*. Conferencia CRESST, septiembre.

van der Linden, W. J., y Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer Verlag.

Wainer, H. and Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education, 6*, 103-118.