Psicothema 2003. Vol. 15, nº 1, pp. 94-100
www.psicothema.com

ISSN 0214 - 9915 CODEN PSOTEG

Análisis de un test mediante el modelo de Rasch

Gerardo Prieto y Ana R. Delgado
Universidad de Salamanca

La denominada Teoría Clásica de los Tests ha sido el principal modelo psicométrico empleado en la construcción y análisis de tests. Sin embargo, sus limitaciones han llevado a la propuesta de modelos alternativos, de los cuales el más parsimonioso es el modelo de Rasch, que permite –dado un buen ajuste de los datos– la medición conjunta de personas e ítems en una misma dimensión o constructo. Ésta y otras ventajas del modelo se presentan siguiendo como ejemplo el análisis del Test de Matemáticas (TM) construido por los autores. El análisis del TM nos ha permitido ilustrar las ventajas del modelo de Rasch tanto en la evaluación colectiva como en el diagnóstico individual, así como presentar las principales técnicas empleadas en el proceso.

Rasch-modelling a Test. Classical Test Theory (CTT) has been the main psychometrical model for constructing and analysing tests. However CTT limitations have given place to alternative models, such as the Rasch Model (RM), which allows –given a good fit– conjoint measurement of persons and items on the same dimension, or construct. The various advantages of the RM are presented following a detailed example– the analysis of the Mathematics Test (MT) constructed by the authors. The MT is used to illustrate the advantages of the RM both for collective assessment and for individual diagnosis; the main techniques used in the analysis are also introduced.

Desde comienzos del siglo XX, la construcción y el uso de tests psicométricos se ha basado principalmente en la Teoría Clásica de los Tests (TCT), un modelo simple, flexible y muy conocido (Gulliksen, 1950), pero que no está exento de limitaciones (Embretson y Hershberger, 1999).

En 1960 el matemático danés Georg Rasch propuso un modelo de medida que permite solventar muchas de las deficiencias de la TCT y construir pruebas más adecuadas y eficientes. El objetivo de este trabajo es exponer las características del modelo, sus ventajas y aplicaciones, mediante la construcción y análisis de una prueba de matemáticas dirigida al segundo curso de la Educación Secundaria Obligatoria (ESO).

El modelo de Rasch

El modelo propuesto por Rasch (1960) se fundamenta en los siguientes supuestos:

1. El atributo que se desea medir puede representarse en una única dimensión en la que se situarían conjuntamente las personas y los ítems.
2. El nivel de la persona en el atributo y la dificultad del ítem determinan la probabilidad de que la respuesta sea correcta. Si el control de la situación es adecuado, esta expectativa es razonable y así debe representarla el modelo matemático elegido. Rasch usó la función logística para modelar la relación:

$$\ln (P_{is} / 1 - P_{is}) = (\theta_s - \beta_i) \quad (1)$$

La ecuación (1) indica que el cociente entre la probabilidad de una respuesta correcta y la probabilidad de una respuesta incorrecta a un ítem (P_{is} /1 - P_{is}), es una función de la diferencia en el atributo entre el nivel de la persona (θ_s) y el nivel del ítem (β_i). Así, cuando una persona responde a un ítem equivalente a su umbral de competencia, tendrá la misma probabilidad de una respuesta correcta y de una respuesta incorrecta (P_{is} /1 - P_{is}= 0,50/0,50). En este caso, el logaritmo natural de P_{is} /1 - P_{is}, refleja que la dificultad del ítem es equivalente al nivel de competencia de la persona (θ_s - β_i= 0). Si la competencia del sujeto es mayor que la requerida por el ítem (θ_s - β_i >0), la probabilidad de una respuesta correcta será mayor que la de una respuesta incorrecta. Por el contrario, si la competencia del sujeto es menor que la requerida por el ítem (θ_s - β_i <0), la probabilidad de una respuesta correcta será menor que la de una respuesta incorrecta,.

Una formulación más conocida del modelo de Rasch, por su difusión en los textos de Teoría de Respuesta a los Ítems (TRI) (Embretson y Reise, 2000; Hambleton, Swaminathan y Rogers, 1991; Muñiz, 1997), se deriva de la predicción de la probabilidad de responder correctamente al ítem a partir de la diferencia en el atributo entre el nivel de la persona (θ_s) y el nivel del ítem (β_i). En este caso,

$$P_{is} = e^{(\theta s - \beta i)} / 1 + e^{(\theta s - \beta i)} \quad (2)$$

Donde e es la base de los logaritmos naturales (2,7183).

Fecha recepción: 16-4-02 • Fecha aceptación: 31-7-02
Correspondencia: Gerardo Prieto Adánez
Facultad de Psicología
Universidad de Salamanca
37005 Salamanca (Spain)
E-mail: gprieto@usal.es

Los valores escalares de las personas y los ítems pueden expresarse en distintas métricas (Embretson y Reise, 2000). La más utilizada es la escala *logit*, que es el logaritmo natural de P_{is} /1 - P_{is}, es decir, $\theta_s - \beta_i$. La localización del punto 0 de la escala es arbitraria. En la tradición de Rasch, se suele situar dicho punto en la dificultad media de los ítems. En este caso, es muy sencilla la interpretación de los parámetros de las personas (los valores de θ_s mayores que 0 indican que las personas tienen una probabilidad superior a 0,50 de éxito en los ítems de dificultad media). Aunque la escala *logit* puede adoptar valores entre más y menos infinito, la gran mayoría de los casos se sitúa en el rango ± 5. Otros usuarios del modelo prefieren, considerando los objetivos y la muestra utilizada, situar el punto 0 en la habilidad media de las personas. Asimismo, la familiaridad con la distribución normal ha llevado a multiplicar por la constante 1,7 el exponente de la ecuación (2) para asimilar la escala *logit* a aquélla. En este caso, la media y la desviación típica de la escala son similares a las de las conocidas puntuaciones típicas z (0 y 1 respectivamente). Por tanto, la casi totalidad de los casos se incluye en el rango ± 3.

Estimación de los parámetros

El objetivo inicial de la administración de un test consiste en estimar los parámetros de los sujetos (θ_s) y de los ítems (β_i) en la variable de interés. En algunas ocasiones, se conoce previamente uno de estos conjuntos de parámetros. Una situación frecuente consiste en estimar los parámetros de las personas a partir de parámetros de ítems ya conocidos (obtenidos en anteriores aplicaciones de la prueba). En este caso, el procedimiento a utilizar sería la *estimación condicional*. Cuando se desconocen los parámetros de ítems y personas, el proceso es denominado *estimación conjunta*. Una descripción detallada de los procedimientos de estimación está fuera del alcance de este artículo. Los lectores interesados pueden encontrarlas en Embretson y Reise (2000), Hambleton, Swaminathan y Rogers (1991) y Muñiz (1997), entre otros. La lógica general del método más usual, denominado de *máxima verosimilitud*, consiste en determinar los parámetros que hacen más probables las respuestas observadas. En el caso de la estimación condicional de los parámetros de las personas, el procedimiento es similar a un proceso de búsqueda: conocidos los parámetros de los ítems, se calcula la probabilidad conjunta de las respuestas observadas a los ítems para cada puntuación θ. Se asigna a cada persona, el valor θ más probable para su patrón de respuestas. Este valor es denominado *estimador de máxima verosimilitud* (θ'). Los procedimientos de cálculo son sumamente largos, por lo que es imprescindible recurrir a programas de ordenador. Algunos de los más utilizados son: Quest (Adams y Khoo, 1996), RASCAL (Assessment Systems Corporation, 1995), RUMM (Sheridan, Andrich y Luo, 1996) y WINSTEPS (Wright y Linacre, 1998).

Los estimadores de θ son asintóticos e insesgados cuando los tests son suficientemente largos. Su desviación típica, denominada *error típico de medida*, es igual a:

$$SE(\theta') = 1 / \sqrt{TI(\theta)} \quad (3)$$

El valor TI(θ) se llama *función de información del test.* Puesto que el error típico de medida es una función inversa de la información del test, este concepto tiene un significado similar al de fiabilidad en la TCT. La función de información del test es igual a la suma de las funciones de información de los ítems que lo integran:

$$TI(\theta) = \sum I(\theta) \quad (4)$$

Donde la función de información del ítem es:

$$I(\theta) = P_i(\theta)\,(1 - P_i(\theta)) \quad (5)$$

De la ecuación (5) se infiere que: (i) la información de un ítem varía a lo largo del continuo y (ii) el punto en el que un ítem aporta la máxima información es el que equivale a su parámetro de dificultad ($\theta_s = \beta_i$).

Ventajas del modelo de Rasch

Las ventajas del modelo de Rasch respecto a la TCT y a otros modelos TRI han sido ampliamente difundidas (Andrich, 1988; Bond y Fox, 2001; Embretson y Hershberger, 1999; Embretson y McCollam, 2000; Embretson y Reise, 2000; Hambleton, Swaminathan y Rogers, 1991; Wright y Stone, 1979). Destacaremos aquí las características que, a nuestro juicio, son más relevantes: medición conjunta, objetividad específica, propiedades de intervalo y especificidad del error típico de medida.

Medición conjunta: Significa que los parámetros de las personas y de los ítems se expresan en las mismas unidades y se localizan en el mismo continuo. En primer lugar, esta propiedad confiere al modelo de Rasch un carácter más realista que el de la TCT, puesto que no es razonable mantener el supuesto de la invarianza de los ítems: es obvio que no todos los ítems miden la misma cantidad del constructo. En segundo lugar, esta característica permite analizar las interacciones entre las personas y los ítems. En consecuencia, la interpretación de las puntuaciones no se fundamenta necesariamente en normas de grupo, sino en la identificación de los ítems que la persona tiene una alta o baja probabilidad de resolver correctamente. Esta característica dota al modelo de Rasch de una gran riqueza diagnóstica.

Objetividad específica: Una medida sólo puede ser considerada válida y generalizable si no depende de las condiciones específicas con que ha sido obtenida. Es decir, la diferencia entre dos personas en un atributo no debe depender de los ítems específicos con los que sea estimada. Igualmente, la diferencia entre dos ítems no debe depender de las personas específicas que se utilicen para cuantificarla. Esta propiedad fue denominada objetividad específica por Rasch (1977).

Supóngase que dos personas de distinto nivel contestan al mismo ítem. De acuerdo con la ecuación (1):

$$\ln (P_{i1} / 1 - P_{i1}) = \theta_1 - \beta_i \text{ , y } \ln (P_{i2} / 1 - P_{i2}) = \theta_2 - \beta_i \text{ .}$$

La diferencia entre ambas personas será igual a:

$$\ln (P_{i1} / 1 - P_{i1}) - \ln (P_{i2} / 1 - P_{i2}) = (\theta_1 - \beta_i) - (\theta_2 - \beta_i) = \theta_1 - \theta_2.$$

De forma similar, si la misma persona contesta a dos ítems de diferente dificultad:

$$\ln (P_{1s} / 1 - P_{1s}) = \theta_s - \beta_1, \text{ y } \ln (P_{2s} / 1 - P_{2s}) = \theta_s - \beta_2 \text{ .}$$

La diferencia en dificultad entre ambos ítems será igual a:

$$\ln (P_{1s} / 1 - P_{1s}) - \ln (P_{2s} / 1 - P_{2s}) = (\theta_s - \beta_1) - (\theta_s - \beta_2) = \beta_1 - \beta_2.$$

En consecuencia, si los datos se ajustan al modelo, las comparaciones entre personas son independientes de los ítems administrados y las estimaciones de los parámetros de los ítems no estarán influenciadas por la distribución de la muestra que se usa para la calibración. Nótese que en la TCT las puntuaciones de las personas dependen de los ítems administrados y la dificultad de los ítems puede variar entre grupos de personas. En la propiedad de objetividad específica se fundamentan aplicaciones psicométricas muy importantes como la equiparación de puntuaciones obtenidas con distintos tests, la construcción de bancos de ítems y los tests adaptados al sujeto.

Propiedades de intervalo: Es importante notar que la interpretación de las diferencias en la escala es la misma a lo largo del atributo medido. Es decir, a diferencias iguales entre un sujeto y un ítem le corresponden probabilidades idénticas de una respuesta correcta. Por ello, la escala *logit* tiene propiedades de intervalo. Por el contrario, en la TCT las puntuaciones son casi siempre ordinales. La métrica intervalar tiene gran importancia, puesto que es una condición necesaria para usar con rigor los análisis paramétricos más frecuentemente empleados en las ciencias sociales (análisis de varianza, regresión, etc) y, además, garantiza la invarianza de las puntuaciones diferenciales a lo largo del continuo (un requisito imprescindible en el análisis del cambio).

Especificidad del error típico de medida: Como han subrayado Embretson y Reise (2000), la objetividad específica no implica que la *precisión* de las estimaciones de los parámetros sea similar en distintos conjuntos de ítems y de personas. Si los ítems son fáciles, se estimarán con más precisión los parámetros de los sujetos de bajo nivel. De forma similar, si los sujetos son de alto nivel, se estimarán con mayor precisión los parámetros de los ítems difíciles. En la TCT, se supone que los tests miden con la misma fiabilidad en todas las regiones de la variable. El modelo de Rasch no asume este supuesto tan poco verosímil. Permite, por el contrario: (i) cuantificar la cantidad de información con la que se mide en cada punto de la dimensión y (ii) seleccionar los ítems que permiten incrementar la información en regiones del atributo previamente especificadas. Este último aspecto es de sumo interés en los tests referidos al criterio, en los que interesa maximizar la fiabilidad en torno a los puntos de corte.

Ajuste de los datos al modelo

Las ventajas del modelo de Rasch sólo pueden ser obtenidas si los datos empíricos se ajustan al modelo. De acuerdo con la ecuación (2), la probabilidad de respuesta a un ítem depende sólo de los niveles de la persona y el ítem en el atributo medido. La presencia de respuestas aberrantes tales como que personas poco competentes resuelvan correctamente ítems difíciles, indicarían que los parámetros de sujetos e ítems son meros numerales carentes de significado teórico. La falta de ajuste podría deberse a diversos factores: multidimensionalidad o sesgo de los ítems, falta de precisión en el enunciado o en las opciones, respuestas al azar, falta de motivación o cooperación, errores al anotar la respuesta, copiado de la solución correcta, etc (Karabatsos, 2000a). Los procedimientos de análisis permiten detectar los ítems y las personas que no se ajustan al modelo. Se han propuesto diversos estadísticos para evaluar el ajuste de los datos (Karabatsos, 2000a, 2000b; Masters y Wright , 1996; Meijer y Sijtsma, 2001; Smith, 2000). Aquí mencionaremos los estadísticos basados en *residuos* (diferencias entre las respuestas observadas y las esperadas), debido a que están implementados en los programas de ordenador más usados. La fórmula de un residuo es:

$$y_{is} = (x_{is} - P_{is}) \quad (6)$$

Donde x_{is} es la respuesta observada y P_{is} la probabilidad de una respuesta correcta de la persona s al ítem i.

Se suelen estandarizar los residuos dividiéndolos por su desviación típica:

$$z_{is} = (x_{is} - P_{is}) / \sqrt{(P_{is} (1 - P_{is}))} \quad (7)$$

Para cuantificar el ajuste al modelo, se emplea preferentemente el estadístico Infit que es la media de los residuos cuadráticos ponderados con su varianza (W_{is}).

$$\text{Infit} = \sum z^2_{is} W_{is} / \sum W_{is} \quad (8)$$

Se puede calcular Infit para un ítem o una persona promediando los valores correspondientes. El valor esperado de este estadístico es 1. Por convención se considera que los valores superiores a 1,3 indican desajuste en muestras con menos de 500 casos, 1,2 en muestras de tamaño medio (entre 500 y 1000 casos) y 1,1 en muestras con más de 1000 casos (Smith, Schumaker y Bush, 1995). Los programas de ordenador aportan representaciones gráficas que facilitan la interpretación de los estadísticos de ajuste.

A continuación, se ilustra la aplicación del modelo de Rasch con el análisis de un test de matemáticas cuyos contenidos corresponden al primer curso de la Enseñanza Secundaria Obligatoria (ESO).

Método

Participantes

Se han analizado los datos de una muestra de 455 alumnos del segundo curso de la ESO (241 varones y 214 mujeres) procedentes de 11 centros públicos de la ciudad de Salamanca. Aunque la muestra no es aleatoria, consideramos que, al tratarse de un curso de educación obligatoria, se han obtenido sujetos a lo largo de todo el continuo de competencia. Se eliminaron los cuestionarios que manifestaban una mala comprensión de las instrucciones o falta de cooperación.

Instrumentos

El test de matemáticas (TM) está compuesto por 30 ítems de elección múltiple organizados en tres bloques de 10 preguntas cada uno. Las preguntas se construyeron a partir de los contenidos de los libros de texto correspondientes al primer curso de la ESO (los estudiados por los participantes en el curso anterior). Los contenidos de los dos primeros bloques son operaciones rutinarias de aritmética y geometría respectivamente; el último está integrado por problemas (5 de aritmética y 5 de geometría).

Cada pregunta se compone de un enunciado y cuatro opciones de las que una es correcta. Las opciones de respuesta fueron ordenadas cuando correspondían a cantidades. En el resto de los casos, fueron aleatorizadas con la condición de que el número fuera similar en cada una de las localizaciones.

Procedimiento

La aplicación del test tuvo lugar durante el mes de marzo de 2001 siguiendo las recomendaciones éticas usuales. El test fue administrado en las clases habituales de cada grupo de alumnos durante las primeras horas de la mañana. En cada grupo se impartieron detalladamente las instrucciones, seguidas por varios ítems de práctica. Se insistió especialmente en que los ítems fueran resueltos mentalmente, usándose el bolígrafo sólo para marcar la respuesta. Una vez impartidas las instrucciones y contestados los ítems de práctica, se informó a los participantes de que disponían de 25 minutos para contestar a la prueba. El tiempo de aplicación fue suficiente, puesto que el 96,26% de los participantes terminó la prueba.

Resultados y discusión

Las respuestas fueron codificadas dicotómicamente y los datos analizados mediante el programa Quest (Adams y Khoo, 1996). En primer lugar, presentaremos los resultados del análisis del ajuste al modelo de los ítems y de los participantes. Como ya se ha comentado, el ajuste es crucial; en su ausencia, los valores carecen de significado teórico y las ventajas del modelo de Rasch se desvanecen. Se han utilizado estadísticos de ajuste global y compro-

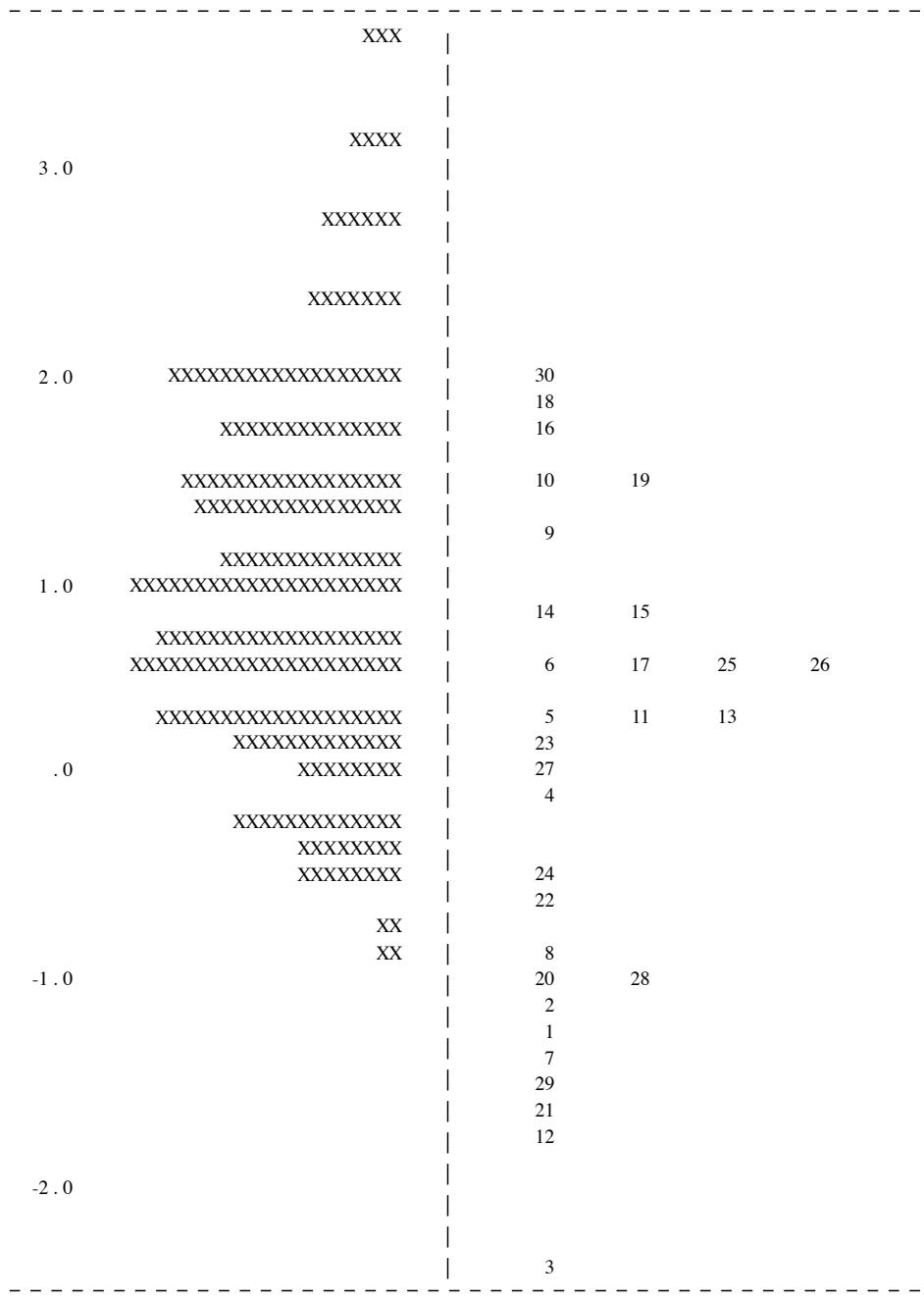

Figura 1. *Escalamiento conjunto de ítems y personas. Cada X representa a dos personas*

baciones de la invarianza de los parámetros. Como indicador del ajuste global, se ha empleado el estadístico Infit (Ítems: media= 1,00; desviación típica= 0,09. Personas: media= 0,99; desviación típica= 0,18). Los datos manifiestan un buen ajuste al modelo: por un lado, las medias y las desviaciones típicas de los valores de Infit son las que cabe esperar cuando no hay divergencias sustanciales entre las predicciones del modelo y los datos empírico; por otro, ninguno de los ítems y sólo un 6% de los alumnos presentan un valor superior a 1,3 (mayor Infit= 1,56).

Como hemos expuesto anteriormente, la objetividad específica es una de las propiedades más importantes del modelo de Rasch. Uno de los principales procedimientos que se recomiendan para analizar el ajuste de los datos al modelo consiste en contrastar empíricamente esta propiedad (Hambleton, Swaminathan y Rogers, 1991). Para analizar la invarianza de los parámetros de los ítems, se dividió aleatoriamente la muestra de alumnos en dos submuestras de similar tamaño, se estimaron los parámetros de dificultad de los ítems en cada submuestra y se llevó a cabo una regresión lineal simple entre los valores β obtenidos. La correlación entre ambos conjuntos de parámetros, la ordenada en el origen y la pendiente de la recta fueron respectivamente 0,973, - 0,001 y 0,985. Los valores esperados en un ajuste perfecto serían 1, 0 y 1 respectivamente. En consecuencia, se puede concluir que se cumple la propiedad de invarianza de los parámetros de los ítems.

La mayor ventaja del modelo de Rasch respecto de la TCT es la propiedad de medición conjunta: los parámetros de ítems y personas están en la misma escala. En la Figura 1, aparece una representación gráfica del escalamiento conjunto. Puede observarse una representación de los objetos (personas e ítems) en un rango de valores entre -3 y 4 *logit*. Las distribuciones de los parámetros de los ítems y de los alumnos son aproximadamente normales, pero no se superponen.

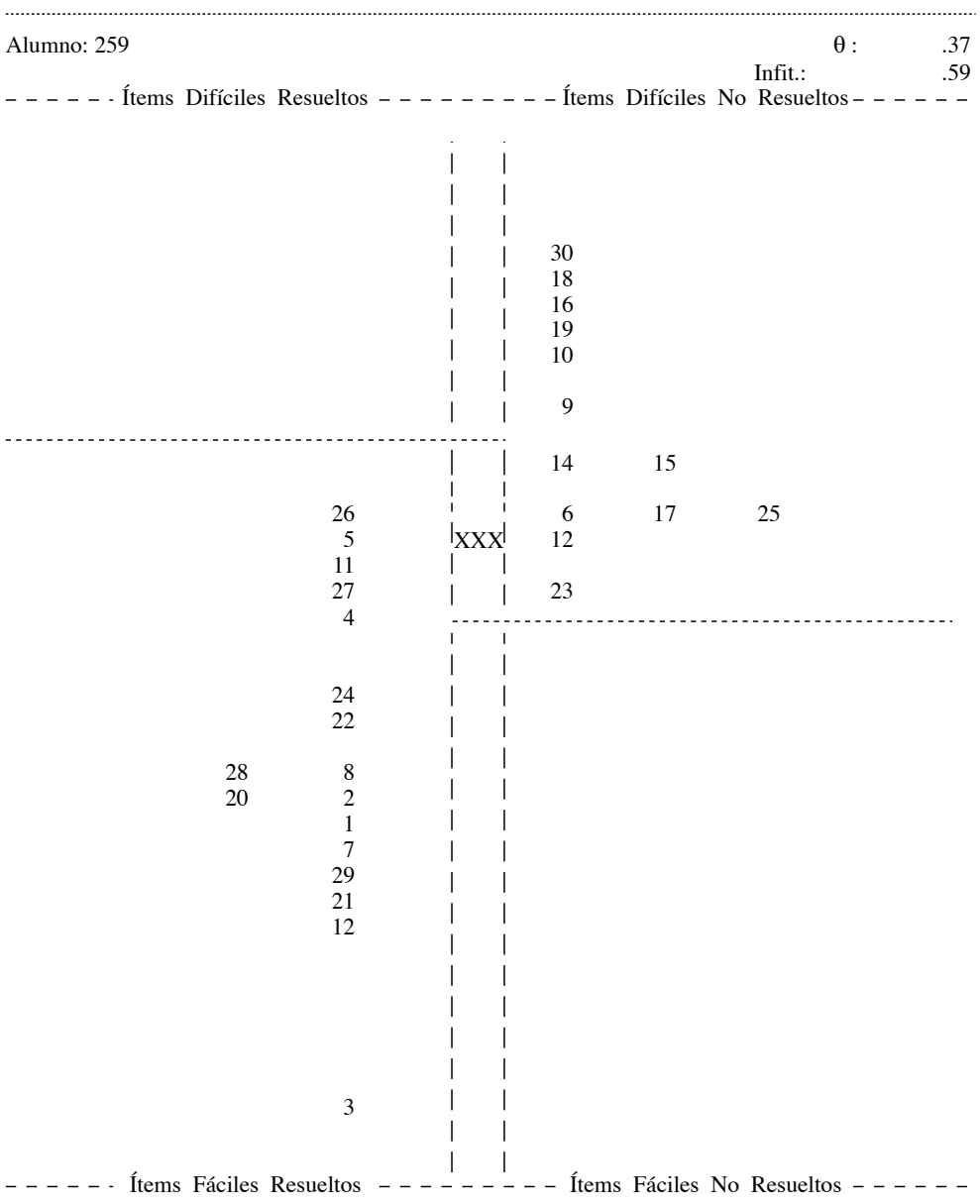

Figura 2. *Mapa de la ejecución de un alumno*

El escalamiento conjunto permite obtener interpretaciones de sumo interés. Mencionaremos brevemente las más importantes.

Nivel del grupo de alumnos en el atributo medido. En el caso de que los ítems fuesen una muestra representativa de los indicadores que permiten indagar acerca de la competencia básica en la comprensión de los conceptos matemáticos y en la resolución de problemas, se podría afirmar que el rendimiento de la muestra es elevado, puesto que la mayor parte de los alumnos tienen puntuaciones superiores a 0 (la dificultad promedio de los ítems). Este dato significa que la mayoría de la muestra tiene una alta probabilidad de resolver correctamente un gran número de ítems. Puesto que el TM se diseñó a partir de los objetivos del curso anterior, este dato coincide con lo esperado.

Adecuación de la prueba al nivel de competencia. En el caso de que el test no tuviese la finalidad de evaluar sólo las habilidades básicas, los datos indicarían que el test es demasiado fácil para la muestra analizada. Como ya hemos indicado, la utilidad de un test para evaluar a los alumnos de forma precisa se incrementa *ajustando* la dificultad de los ítems al nivel de competencia. Por tanto, se observa que faltan ítems de alta dificultad (ítems con $\beta > 2$) que serían más apropiados para evaluar adecuadamente a los sujetos con alta competencia. Es decir, la representación conjunta facilita la identificación de regiones del continuo que no han sido suficientemente muestreadas.

Definición del constructo. En ocasiones, la finalidad prioritaria del escalamiento no es escalar sujetos, sino indicadores de un constructo. El objetivo puede ser responder a preguntas tales como: ¿los indicadores del constructo se pueden escalar en una sola dimensión?, ¿cuál es la diferencia en el continuo de competencia entre conjuntos de ítems que comparten ciertas características (por ejemplo, operaciones aritméticas y geométricas básicas: suma de números enteros y cálculos con rectas en el plano)?, ¿cuáles son las características de las tareas representativas de la alta competencia? Con fines ilustrativos, contestaremos a estas preguntas con los datos obtenidos con el TM. Por un lado, el ajuste de los datos apoya la unidimensionalidad del test. Por otro, los promedios de los valores de los ítems de sumas y cálculos con rectas son -1,13 y 1,25 respectivamente; en consecuencia, la resolución de este tipo de contenidos requiere muy distinto nivel de competencia.

La indagación acerca de las características de los ítems representativos de los distintos niveles de competencia resulta muy útil para dotar de significación al constructo medido. Por ejemplo, en el test TM los ítems 30, 18 y 16 son los más difíciles (Véase la Figura 1). Estos ítems corresponden a tareas geométricas que requieren la integración de cálculos y representaciones espaciales de cierta complejidad.

La formulación de modelos para explicar la dificultad de los ítems a partir de los procesos mentales y las estructuras de conocimiento requeridas por la tarea es una de las extensiones del modelo de Rasch más prometedoras. Desde el enfoque representacional (Embretson, 1983), se ha propuesto que los procedimientos de validación del constructo no deben fundamentarse sólo en las correlaciones con criterios, sino en la explicación de las variaciones intratarea: la dificultad del ítem se considera como un indicador de la *complejidad cognitiva* requerida para resolverlo correctamente (Prieto y Delgado, 1999, 2000). La complejidad cognitiva se explica por los procesos, las estrategias y las estructuras de conocimiento subyacentes a la ejecución del ítem. Las extensiones del modelo de Rasch propuestas por Fischer (1973) y Embretson (1997), entre otros, tienen esta finalidad.

Diagnóstico individual. La representación gráfica conjunta se puede llevar a cabo a nivel individual, de forma que se pueden identificar los ítems que la persona tiene una alta o baja probabilidad de resolver correctamente. Desde esta perspectiva, la interpretación de la puntuación de un sujeto es más rica que la simple clasificación mediante baremos o normas de grupo.

Por ejemplo, en la Figura 2 aparece un mapa de la ejecución de un sujeto de nivel medio.

Este mapa representa conjuntamente el nivel del sujeto (XXX) y el de los ítems en el continuo. Las dos líneas de puntos representan un intervalo de $\theta \pm SE(\theta')$. Los ítems que el alumno tiene una baja probabilidad de resolver correctamente son los situados sobre la línea de puntos en el lado derecho del mapa. Los ítems que el alumno tiene una alta probabilidad de resolver correctamente son los situados bajo la línea de puntos en el lado izquierdo.

Los mapas son también muy útiles para interpretar los patrones de respuestas aberrantes. Si el patrón de respuestas de la persona se ajusta, se espera que la mayor parte de los ítems situados bajo el nivel del sujeto se sitúe en la parte inferior izquierda del gráfico (bajo la línea de puntos) y que la mayor parte de los ítems situados sobre el nivel del sujeto se sitúe en la parte superior derecha (sobre la línea de puntos). En este caso, el ajuste al modelo sería muy elevado. La presencia de ítems en los cuadrantes superior izquierdo e inferior derecho reflejan respuestas inesperadas.

Anteriormente hemos señalado que el supuesto de invarianza del error típico de medida asumido por la TCT es poco verosímil. Que se mida con menor precisión en los extremos del continuo es la situación mas frecuente. El modelo de Rasch permite estimar específicamente la fiabilidad de cada medida, mediante el error típico de medida (fórmula 3) o la función de información (fórmula 4). En la Figura 3, se representan los errores típicos de medida en los distintos niveles de la variable.

Puede observarse que el test mide con mayor precisión en el rango central de la escala. Por ejemplo, los intervalos de estimación del valor paramétrico a un nivel de confianza del 95% difieren notablemente en el centro y en los extremos. Para $\theta' = 4$, el intervalo es de $\pm$ 2 logit, mientras que para $\theta' = 0$ es de $\pm$ 0,8 logit. En caso de que se deseara incrementar la precisión en los niveles altos de competencia, habría que incluir en el test ítems de elevada dificultad (obsérvese que en el TM no existen ítems con $\beta > 2$). Así, las distribuciones de los valores de la función de información

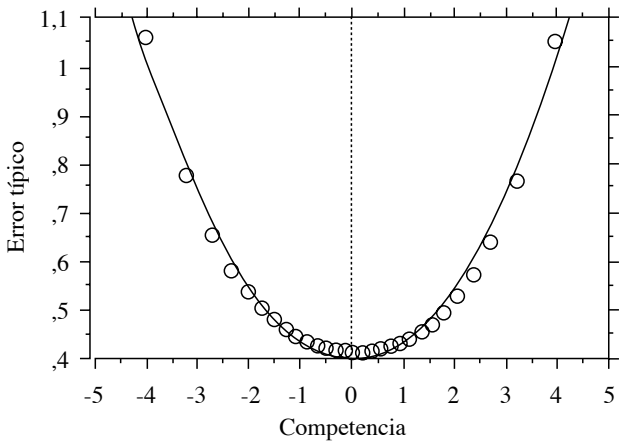

Figura 3. *Errores típicos de medida en los distintos niveles de competencia*

o del error típico de medida resultan especialmente útiles para identificar las regiones del continuo en las que se mide de forma poco precisa. Además, suelen servir como criterio para construir tests a partir de bancos de ítems: puesto que la función de información del test es la suma de las funciones de información de los ítems, es posible seleccionar aquéllos que permitan medir con mayor precisión en un rango determinado.

En conclusión, el análisis del TM mediante el modelo de Rasch nos ha permitido ilustrar algunas de las ventajas de éste tanto en la evaluación colectiva como en el diagnóstico individual, así como presentar las principales técnicas empleadas en el proceso.

Nota

Esta investigación ha sido financiada por la Dirección General de Investigación del Ministerio de Ciencia y Tecnología (Departamento Técnico de Promoción General del Conocimiento. Nº del Proyecto: PB98-0263).

Referencias

Adams, R.J. y Khoo, S. (1996). *Quest: The interactive test analysis system*. Victoria: ACER.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park: Sage.

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

Bond, T.G. y Fox, C.M. (2001). *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah, NJ: LEA.

Embretson, S.E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 52,* 179-197.

Embretson, S. E. (1997). Multicomponent response models. En W.J. van der Linden y R.K. Hambleton (Eds.), *Handbook of modern item response theory.* (pp. 305-321). New York: Springer.

Embretson, S.E. y Hershberger, S.L. (1999). *The new rules of measurement*. Mahwah, NJ: LEA.

Embretson, S.E. y McCollam, K.M.S. (2000). Psychometric approaches to understanding and measuring intelligence. En R.J. Sternberg (De.). *Handbook of intelligence* (pp. 423-444). Cambridge,UK: Cambridge University Press.

Embretson, S.E. y Reise, S.P. (2000) *Item response theory for psychologists.* Mahwah, NJ: LEA.

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

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

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

Karabatsos, G. (2000a). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1*, 152-176.

Karabatsos, G. (2000b). Using Rasch measures for Rasch model fit analysis. *Popular Measurement, 3*, 70-71.

Masters, G.N. y Wright, B.D. (1996). The partial credit model. En W.J. van der Linden y R.K. Hambleton (Eds.). *Handbook of modern item response theory*. New York: Springer.

Meijer, R.R. y Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

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

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.* (207-226) Madrid: Pirámide.

Prieto, G. y Delgado, A.R. (2000) Utilidad y representación en la psicometría actual. *Metodología de las Ciencias del Comportamiento, 2(2),* 111-127.

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

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. En M. Glegvad (De.). *The Danish Yearbook of Philosophy* (pp. 59-94). Copenhagen: Munksgarrd.

Sheridan, B., Andrich, D. y Luo, G. (1996). *Welcome to RUMM: A windows-based item analysis program employing Rasch unidimensional measurement models*. User´s Guide.

Smith, R.M. (2000). Fit Analysis in latent trait measurement models. *Journal of Applied Measurement, 1*, 199-218.

Smith, R.M., Schumaker, R.E. y Bush, M.J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of outcome measurement, 2,* 66-78.

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

Wright, B.D. y Stone, M.H. (1979). *Best test design. Rasch measurement*. Chicago: MESA Press.