Psicothema
2001. Vol. 13, nº 2, pp. 302-309

ISSN 0214 - 9915 CODEN PSOTEG

Medida de la integración del conocimiento estadístico y psicológico

Salvador Algarabel y Eva Rosa
Universidad de Valencia

En un estudio empírico llevado a cabo en condiciones de clase, tres grupos de universitarios completaron ejercicios de diseño experimental realizados sobre descripciones de investigaciones con contenido psicológico, ejercicios abstractos o ejercicios de cálculo. Los tres grupos recibían la misma clase teórica y eran esencialmente iguales. En este contexto se intentó una comparación entre el escalamiento del conocimiento obtenido por análisis de trayectorias (pathfinder) o por escalamiento multidimensional contra una aplicación nueva consistente en aplicar el modelo politómico de crédito parcial generalizado. Los resultados mostraron que este último era claramente superior a los dos primeros, detectando diferencias significativas antes y después del examen de la asignatura. Los resultados se discuten en relación con la posibilidad de transferencia del conocimiento entre dominios.

Measurement of statistical and psychological knowledge integration. In an empirical study carried out in a Psychology experimental design class, three groups of university students completed three types of exercises: calculus, abstract, or exercises on real experimental descriptions. The three groups received the same theoretical lectures and were matched in every respect. In this context we attempted a comparison between multidimensional scaling and pathfinder on the one hand, and a new application of the generalized partial credit model. The result showed that the generalized partial credit model was the best mean to detect significant differences before and after the academic training. The results are discussed in terms of the possibility of transfer among different knowledge domains.

La adquisición de conocimiento científico requiere la acción de un conjunto de procesos complejos por parte del estudiante que da lugar a un conjunto integrado de conceptos, usables en la solución de problemas nuevos. El estudio del diseño experimental es una de estas situaciones de aprendizaje complejo. En este contexto, el estudiante debe comprender la descripción psicológica de un problema y la lógica del análisis de datos. Cuando ambos tipos de conceptos; los psicológicos y los estadísticos se dominan, se puede llegar a juzgar la adecuación de un diseño para una situación psicológica específica o diseñar un experimento nuevo para responder a una pregunta científica. En una situación como esta, la estructura relacional, estática o dinámica, del conjunto de conceptos adquiere importancia primordial para determinar «lo que se sabe», y nos refiere a lo que habitualmente se entiende como «conocimiento experto».

Muchos tipos distintos de estructuras teóricas se han creado para intentar capturar los aspectos importantes de este conocimiento experto. Redes semánticas (Collins y Loftus, 1975), sistemas de producción (Anderson, 1983), scripts (guiones), esquemas y modelos mentales o situacionales (Zwaan y Radvansky, 1998) son algunos de ellos. Cuando este problema de investigación se analiza en el campo del dominio de conocimiento académico, no siempre está claro a qué se debe el rendimiento superior del experto en un dominio de conocimiento. Podría deberse a la disposición de un conocimiento estratégico o a la superior estructuración del dominio, o a una combinación de ambos. Estudios realizados sobre dominios complejos, tales como la física (Zajchowski y Martin, 1993), programación (Cañas, Bajo, y Gonzalvo, 1987; Pennington, 1987), o la enseñanza de las matemáticas (Gómez, Hadfield, y Housner, 1996) revelan la gran importancia de la estructuración del conocimiento. De todas formas, existe evidencia amplia en solución de problemas (por ej Schoenfeld, 1989) que revelan claramente que el conocimiento estratégico es en muchas ocasiones de importancia primordial y muchas veces el experto apela a él para buscar respuestas en dominios poco familiares o a dar respuestas muy eficientes en aquellos que sí lo son.

Si bien se han desarrollado muchos puntos de vista teóricos distintos para captar la complejidad del conocimiento científico, no existen tantas técnicas cuantitativas para captarlo (por ej. véase Reitman y Biolsi, 1991 para una revisión de muchas de las técnicas de naturaleza más bien cualitativa). Algunas veces, los investigadores demuestran el nivel de conocimiento complejo de un estudiante o un experto por su capacidad para resolver preguntas difíciles, o a través del análisis de los protocolos de respuesta. Sin embargo, lo conveniente es tener una medida del nivel de complejidad global adquirido en el dominio. Este nivel de complejidad requeriría un índice o un conjunto completo de índices que reflejara de alguna forma la estructura dinámica de conceptos expresados de forma general o específica, siendo este aspecto función de la situación de prueba. El escalamiento multidimensional (por ej. Young y Hamer, 1987; también la aglomeración jerárquica, John-

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

son, 1967, y los árboles aditivos, Sattath y Tversky, 1977) se ha utilizado con este propósito (en aplicaciones psicológicas: Adelson, 1981; Rips, Shoben y Smith, 1973). Si el conocimiento avanzado se organiza en términos de dimensiones relacionales abstractas, entonces esta técnica es capaz de extraer información de los juicios relacionales y situarlos en relación con un conjunto de dimensiones inferidas. Existen muchos datos empíricos (por ej. Zajchowski y Martin, 1993) que indican que el avance en un campo de conocimiento se caracteriza por el desarrollo de principios organizaciones generales. Por ello, las dimensiones que definen el espacio sobre el que están organizados los conceptos representan algún tipo de propiedad general, reflejando las distancias espaciales entre los conceptos su distancia conceptual, y siendo la contrapartida cuantitativa de los datos empíricos frecuentemente obtenidos en la investigación.

Existen otras técnicas más especializadas, y no de tan amplio uso, también basadas en el análisis de los datos de proximidad, tales como el análisis de trayectorias (pathfinder, Goldsmith, Johnson y Acton, 1991; Schvaneveldt, 1990; Schnaveldt, Durso and Dearholt, 1985; véase también una polémica sobre sus fundamentos en Arabie, 1993, 1994; Schvaneveldt, Cooke, Dearholt, Durso y McDonald, 1994, y la prueba de un teorema de importancia en Douglas Carroll, 1995). Aquí, los conceptos son representados como nodos y sus relaciones por medio de conexiones ponderadas. El análisis de trayectorias se ha aplicado repetidamente en el contexto del estudio de la adquisición de conocimiento experto (Cañas, Bajo, y Gonzalvo, 1994; Gillan, Breedin and Cooke, 1992; Gómez, Hadfield, y Housner, 1996; Rowe, Cooke, Hall and Halgren, 1996; Wilson, 1994). A destacar el hecho de que esta técnica produce resultados intuitivamente acordes con nuestro sentido común; es decir, redes gráficas (e índices estadísticos correspondientes) en donde los conceptos se sitúan relacional y espacialmente.

Tanto el escalamiento multidimensional, como el análisis de trayectorias comparten una serie de conceptos comunes, si bien, el escalamiento multidimensional representa más las relaciones de tipo general, mientras que el «pathfinder» (Gonzalvo, Cañas, y Bajo, 1994; Johnson, Goldsmith, y Teague, 1994) está más orientado a la representación de información local, por lo que el uso de uno u otro puede depender de los objetivos concretos de la investigación.

Es importante notar que, a diferencia de la forma en que se mide el conocimiento complejo en el campo psicométrico por medio de test, en el campo experimental, el conocimiento complejo se sondea preguntando directamente al examinado sobre el nivel de relacionalidad entre los conceptos cuya estructura queremos investigar. Y es este aspecto, el que se tiene muy especialmente en cuenta en los movimientos educativos progresivos contemporáneos para medir el conocimiento complejo de forma cualitativa. Lo que los pedagogos o educadores llaman «mapas conceptuales», no es sino la especificación a priori de una serie de conceptos sobre los que se pide un juicio a veces de diverso tipo, y no meramente juicios de fuerza, tal como se hace en el caso de las técnicas cuantitativas anteriores (por ej., Novak, 1990).

La investigación que se describe a continuación tiene como objetivo la comparación de diversas técnicas para medir el conocimiento complejo, en un contexto en el que se llevan a cabo manipulaciones experimentales en ambientes de aprendizaje naturales. Esto es, mientras que el escalamiento multidimensional es una técnica de escalamiento global, el «pathfinder» es más sensible y representa mejor la información de carácter más local (Cañas et al., 1994). Ambas tienen diversos inconvenientes. Desde el punto de vista práctico, si un dominio está formado por un conjunto numeroso de conceptos, el número de apareamientos totales a realizar y requerido por estas técnicas puede aumentar substancialmente hasta unos niveles impracticables. Desde el punto de vista psicométrico podrían existir modelos mucho más sofisticados a la hora de establecer la relación entre la adquisición y la formulación explícita de algunos parámetros asociados con los ítems.

Nuestro propósito es recoger las ideas educativas sobre medición, avanzadas anteriormente por los movimientos educativos contemporáneos, en el contexto de un sistema riguroso y conveniente de medición. Puesto que el conocimiento avanzado posibilita el establecimiento de relaciones entre conceptos, nuestro propósito es aplicar la teoría de respuesta al ítem a una situación en la que se exige una respuesta polítoma por parte del sujeto. Esta respuesta polítoma pretende reflejar el mismo tipo de información que se intenta capturar por medio de un mapa de conceptos o por medio de un juicio relacional en los procedimientos anteriormente vistos. Esto es, en vez de pedir una elección entre varias alternativas, nuestro objetivo es pedir un juicio de relacionalidad entre dos conceptos sobre una escala tipo Likert, para posteriormente utilizar este juicio como índice del conocimiento. Pero, a diferencia de las técnicas anteriores, pretendemos demostrar adicionalmente, que la aplicación de un modelo de teoría de respuesta al ítem es más sensible y riguroso que los procedimientos tradicionales de medición.

La técnica se aplica a una situación de aprendizaje complejo en el que un grupo de estudiantes universitarios está intentando adquirir los conocimientos elementales del diseño experimental. El diseño experimental es un dominio de conocimiento y un lenguaje instrumental aplicable a muchos dominios distintos de contenido, particularmente la prueba de hipótesis psicológicas. Cuando hay que evaluar un artículo científico, la primera labor del lector es comprender la dinámica conceptual de los términos que se exponen. Puesto que la situación descrita en una investigación se aparta de la experiencia cotidiana, el investigador tiene que construir un modelo mental (o situacional) de las relaciones descritas. Ahora bien, una parte importante de esa relacionalidad conceptual hace referencia al estatus estadístico de cada concepto. Por ejemplo, en la descripción de un problema de investigación alguna variable psicológica descrita puede contribuir a la variabilidad, y, a su vez, la variabilidad, afecta a la potencia, y así sucesivamente. Así pues, la evaluación de una investigación psicológica requiere la integración de un conocimiento psicológico con uno estadístico. También podríamos ver esta situación en términos de analogías (por ej. Gentner y Markman, 1997; Holyoak y Thagard, 1997; Kolodner, 1997). Desde este punto de vista, la descripción de una investigación es el análogo de interés, mientras que el conocimiento estadístico de base constituye el análogo fuente. Para que esta proyección sea útil debe llevarse a cabo una proceso de inferencia basado en las comunalidad compartida entre ambos dominios.

No existen muchas investigaciones en donde se analice concretamente el conocimiento del diseño experimental aplicado a dominios psicológicos. Schraagen (1993) estudió la conducta de expertos e inexpertos en la solución de problemas de diseño experimental en dominios de conocimiento familiares o no familiares. Su estudió detectó que aunque los procedimientos expertos de análisis se exportan a los dominios de conocimiento no familiares, el conocimiento específico del dominio es esencial para llevar a cabo soluciones «buenas» a los problemas de diseño de investigación.

En resumen, tres grupos de estudiantes universitarios que están tomando un curso teórico sobre diseño experimental, reciben y resuelven tres tipos distintos de ejercicios en una clase práctica. Al finalizar el curso, se evalúa su estructura dinámica de conocimiento pidiéndoseles que enjuicien la relacionalidad de una serie de pares de conceptos (estadísticos y psicológicos) referidos a un artículo publicado de investigación. Posteriormente, y dado que este es un estudio naturalista, se vuelve a realizar la evaluación anterior una vez sometidos al examen de la asignatura. Dado los hábitos de estudio de las facultades de psicología, se pretende medir la estructura de conocimiento con seguridad absoluta de que ha habido un estudio intenso, lo que sólo puede garantizarse una vez superado el examen. En este contexto se comparan tres técnicas de escalamiento del conocimiento complejo; dos de ellas ampliamente utilizadas en psicología experimental: «pathfinder» y escalamiento multidimensional, y una aplicación nueva que aquí proponemos, el escalamiento por medio de un modelo de crédito parcial generalizado (Muraki, 1997), modelo típico que sigue los principios de la teoría de respuesta al ítem.

Método

Participantes

Participaron 105 estudiantes de 2º curso de la Facultad de Psicología de Valencia, pertenecientes a tres grupos del módulo práctico «Diseños de Investigación Experimental», integrados por 35, 33 y 37 alumnos, respectivamente. Todos ellos eran alumnos del módulo teórico de la misma asignatura con un profesor único. Posteriormente, a la hora de realizar los cálculos pertinentes diversos sujetos fueron eliminados basándose en distintos criterios que serán expuestos a medida que se presenten los resultados. Los estudiantes completaron todos los ejercicios que se describirán posteriormente como parte de sus obligaciones académicas para superar la asignatura. Cada uno de los tres grupos fue asignado al azar a una de las tres condiciones de tratamiento del diseño: ejercicios psicológicos, ejercicios teóricos y ejercicios prácticos. Se intentó determinar si existía alguna razón por la que los grupos pudieran diferir entre sí, y no pudo encontrarse ninguna. Los grupos tampoco difirieron en la puntuación del examen teórico de la asignatura realizado a final de curso.

Materiales y Procedimiento

Se llevaron a cabo 12 sesiones de enseñanza, una por semana. Durante las dos primeras se llevaron a cabo ajustes de procedimiento para determinar parámetros como la duración de los ejercicios, su número, y permitir un avance suficiente de las clases teóricas. Estas sesiones tuvieron además una función de práctica cara a las sesiones posteriores. A partir de la tercera sesión, los tres grupos recibieron 3 tipos diferentes de ejercicios. Los ejercicios hacían referencia a problemas de diseño experimental, y estaban basados, al igual que las explicaciones teóricas del profesor en el enfoque llamado de comparación de modelos (Maxwell y Delaney, 1990). Estos ejercicios se llamaran a partir de ahora: teóricos, prácticos y psicológicos. En el grupo teórico, los participantes tenían que completar ejercicios de tipo abstracto sobre los conceptos explicados. En la condición práctica, los participantes tenían que resolver ejercicios numéricos, y en la condición psicológica, se realizaban ejercicios sobre la base descripciones de investigaciones empíricas de tipo psicológico. Los ejercicios fueron los mismos en las tres condiciones excepto por las características distintivas señaladas anteriormente. Las sesiones 11 y 12 se dedicaron a evaluar el conocimiento adquirido, si bien la segunda sesión fue inesperada para los estudiantes. La evaluación se llevó a cabo de la siguiente manera. Se dio un artículo publicado, y tomándolo como referencia, los estudiantes tenían que emitir un juicio de relacionalidad entre conceptos en una escala de 1 a 7. Once de los conceptos provenían del artículo, mientras que otros 16 eran conceptos de diseño experimental, y se presentaron todos los pares posibles. Durante la última sesión de prueba, estos 16 conceptos fueron reducidos a 12. Como ejemplo, se pidió al estudiante que indicara qué relación había entre el concepto de «hipótesis nula» y «grado de ansiedad en el examen» (una variable del artículo de investigación). Se dijo a los estudiantes que los conceptos podían relacionarse por medio de muy diversas dimensiones, aunque se esperaba que los evaluaran sobre la base del artículo de investigación. Se les dio un ejemplo. Se aparearon todos los conceptos y se presentaron aleatoriamente en un cuaderno, que también incluía una hoja separada con el listado de todos los conceptos dados (véase el apéndice 1) para que los tuvieran como referencia. Antes de la realización de las sesiones y de la recogida de datos, el profesor de la asignatura respondió a la misma serie de apareamientos, siendo utilizada su respuesta como criterio-experto en relación con el cual el cual se evaluó el aprendizaje.

Resultados

Análisis por medio del modelo de crédito parcial generalizado (Muraki, 1997)

Por razones técnicas (variaciones previsibles en las capacidades discriminativas de los ítems) se ha elegido para escalar los resultados el modelo de crédito parcial generalizado (Muraki, 1997). El modelo de crédito parcial generalizado, una de cuyas formas de expresar la probabilidad de respuesta, es:

$$\frac{\exp\left[\sum_{k=1}^{h}\left[a_i\left(\theta - b_i + d_h\right)\right]\right]}{\sum_{c=1}^{m_i}\exp\left[\sum_{k=1}^{c}\left[a_i\left(\theta - b_i + d_h\right)\right]\right]}$$

es miembro de la familia de modelos de teoría de respuesta al ítem que permite variaciones en discriminación y dificultad. La función anterior permite puntuar escalas politómas de respuesta (k categorías), y en donde a, b, d son los parámetros de discriminación, dificultad (posición del ítem), y frontera categorial e indica que la probabilidad de responder en una categoría de respuesta k, se expresa como la probabilidad condicional sobre las categorías (k-1) y k. El modelo itera para el cálculo de las probabilidades sobre categorías adyacentes

Análisis de Ítems y Estimación de parámetros

No todos los ítems posibles de una escala, en este caso los 351 pares originales, conllevan información de igual entidad acerca de la estructura de conocimiento. Por ello, previo a cualquier análisis se debe ser capaz de utilizar algún criterio de selección. Además, de alguna forma, hemos de tener en cuenta que tan importante es

reconocer la existencia de una relación como la ausencia de la misma entre dos conceptos. Con estas ideas en mente procedió a llevar a cabo el análisis de ítems correspondiente a los modelos TRI.

Para el caso de la escala previa al examen, ésta estaba compuesta de 351 ítems, mientras que la posterior estaba formada por 231. Sin embargo, como la mayor parte de los conceptos de la primera escala eran idénticos a los de la escala 2, se unificaron en un conjunto de ítems comunes, lo que dio lugar a 210 ítems. Estos 210 ítems eran el resultado de aparear 21 conceptos entre sí. Esta matriz con evaluaciones politómicas entre 1 (poco relacionado) y 7 (máximamente relacionado) constituyeron la matriz de partida de todos los análisis que a continuación se describen.

El juicio de similaridad emitido por cada examinado fue referido al criterio del profesor de la asignatura, restando de 7 el valor absoluto de la diferencia, con lo que la relacionalidad se convirtió en una escala parcial de «corrección» de 1 a 7 puntos. La resultante es una evaluación parcial del conocimiento hecho de forma objetiva, rápida y estandarizada. Si no existiese una referencia al criterio, entonces la resultante implicaría simplemente el criterio del alumno que puede o no reflejar el conocimiento adecuado. Se analizó la distribución de respuesta de los participantes en función de las categorías de respuesta, observándose un mayor número de respuestas hacia la parte superior de la escala que a la parte inferior. Este desequilibrio de respuesta produce con seguridad problemas graves de estimación cuando se calibren los datos por cualquiera de los procedimientos que se van a introducir a continuación. Para evitarlo, se agruparon las tres categorías inferiores, dando lugar a una escala de 1 a 5, con la distribución de respuestas bastante uniforme.

A continuación se calcularon las correlaciones poliseriales de la respuesta a cada ítem con la puntuación total obtenida por cada participante (índice de discriminación). Este cálculo se realizó separando los ítems en tres subescalas: la escala mixta, compuesta de 108 ítems en donde uno de ellos era estadístico, y el otro psicológico; la escala psicológica, compuesta de 36 ítems, en donde ambos conceptos provenían del artículo de lectura, y la escala estadística, en donde los 66 pares que la componían eran estadísticos. Se eliminaron todos los ítems con una correlación inferior a 0.30. El análisis resultante produjo tres subescalas, a las que desde ahora nos referiremos como escala mixta (50 ítems), escala psicológica (14 ítems), y escala estadística (22 ítems). Cada una de estas subescalas fue sometida separadamente a un análisis factorial de información completa (Wilson, Wood, y Gibbons, 1998) para determinar la dimensionalidad de las escalas correspondientes. Puesto que la escala mixta era claramente multidimensional y la estadística presentaba signos de multidimensionalidad se procedió a realizar una selección adicional de ítems para conseguir la unidimensionalidad requerida por el modelo de teoría de respuesta al ítem que va a ser utilizado para escalar los datos. Sólo se retuvieron los ítems que saturaban en el primer factor, lo que condujo a tres subescalas con 16, 10, y 11 ítems, correspondiendo respectivamente a la escala mixta, la escala psicológica, y la escala estadística (véase apéndice II).

Análisis estadísticos sobre las estimaciones de los sujetos

La habilidad de los sujetos fue estimada de acuerdo con el procedimiento de máxima verosimilitud ponderada (Warm, 1989; todas las estimaciones sobre la habilidad de los examina-

Apéndice I
Listado completo de conceptos evaluados en el estudio

sesion	Conceptos estadísticos	Conceptos psicológicos (artículo)
1	Modelo completo	Las ayudas mejoran el aprendizaje
	Modelo restringido	Los estudiantes universitarios se beneficiarán más
	Error completo	Los estudiantes universitarios se beneficiaran más que los de bachillerato
	Error restringido	Puntuación en preguntas aplicadas
	potencia	Puntuaciones sobre el reconocimiento
	alfa por contraste	Puntuaciones en preguntas de verdadero-falso
	alfa por experimento	Tipo de ayuda
	F	Tipo de instrucción
	Efecto principal	Tipo de información
	Efecto de interacción	Conocimiento previo del texto
	Tamaño del efecto	Nivel de ansiedad en el examen
	Comparaciones a posteriori	
	Diseño factorial	
	Puntuación individual	
	Hipótesis nula	
	Hipótesis alternativa	
2	Modelo completo	Las ayudas mejoran el aprendizaje
	Modelo restringido	Puntuación en preguntas aplicadas
	Error completo	Puntuaciones sobre el reconocimiento
	Error restringido	Puntuaciones en preguntas de verdadero-falso
	Hipótesis nula	Tipo de ayuda
	Hipótesis alternativa	Tipo de relación
	F	Tipo de información
	Efecto principal	Conocimiento previo del texto
	Efecto de interacción	Nivel de ansiedad en el examen
	Tamaño del efecto	Sexo de los sujetos experimentales
	Comparaciones a posteriori	

dos fueron realizadas por PARSCALE, Muraki y Bock, 1997) de acuerdo con el modelo de crédito parcial generalizado, anteriormente brevemente descrito. Las estimaciones resultantes fueron sometidas a un análisis mixto de varianza de 3 (grupo) x 2 (fase) x 3 (escala), en el que el segundo y tercer factores fueron intrasujeto.

La fase fue significativa [$F(1,67)= 226,65$, $p<0.01$, $MCe= 1.098$], indicando que las estimaciones para la segunda fase dieron valores considerablemente superiores a la de la primera. La escala también fue significativa [$F(2,134)= 33.54$, $p<0.01$, $MCe= 1.443$], así como la interacción de fase por escala [$F(2,134)= 95.61$, $p<0.01$, $MCe= 0.573$]. Los valores medios para el tipo de escala fueron 0.7673, 0.4594 y 1.1528 para la escala mixta, psicológica y estadística respectivamente.

Las comparaciones a posteriori mostraron que todas los contrastes fueron significativos al nivel de $p= 0.01$, excepto la diferencia entre las escalas mixta y psicológica que lo fue al 0.05. Los análisis de efectos simples mostraron que el aumento en conocimiento entre la fase 1 y 2 fue mayor para la escala estadística [$F(1,67)= 67.15$, $MCe= 0.99$] y menor para la escala psicológica [$F(1,67)= 18.95$, $MCe= 0.39$]. La variable grupo fue marginalmente significativa [$F(2,67)=12.39$, $p=0.09$, $MCe=2.508$] con estimaciones media de 0.9947, 0.5736 y 0.8111 para los grupos psicológico, teórico y práctico respectivamente. Ni las interacciones de grupo por fase [$F(2,67)= 0.71$, $p= 0.73$, $MCe= 1,098$], ni grupo por escala [$F(4,134)= 10.07$, $p=.14$, $MCe= 1.443$], ni la interacción de orden superior de grupo por fase y escala fueron significativos [$F(4,134)= 1.49$, $p= 0.63$, $MCe= 0.53$].

Apéndice II
Escalas finales después del análisis de ítems

Lista completa de ítems

I	Concepto 1	Concepto 2
1. Escala Mixta		
1	Las ayudas mejoran el aprendizaje	Comparaciones a posteriori
2	Modelo restringido	Nivel de ansiedad en el examen
3	Error completo	Tipo de relación
4	Puntuación en preguntas aplicadas	Efecto principal
5	Puntuación en reconocimiento	Potencia
6	Puntuación en reconocimiento	Estadístico F
7	Potencia	Puntuación en preguntas de verdadero-falso
8	Potencia	Tipo de relación
9	Potencia	Conocimiento previo del texto
10	Puntuación en preguntas de verdadero-falso	Hipótesis alternativa
11	Puntuación en preguntas de verdadero-falso	Tamaño del efecto
12	Hipótesis alternativa	Nivel de ansiedad en el examen
13	Tipo de relación	Comparaciones a posteriori
14	Hipótesis nula	Conocimiento previo del texto
15	Estadístico F	Nivel de ansiedad en el examen
16	Tipo de información	Comparaciones a posteriori
2. Escala Psicológica		
1	Las ayudas mejoran el aprendizaje	Nivel de ansiedad en el examen
2	Puntuación en preguntas aplicadas	Nivel de ansiedad en el examen
3	Puntuación en reconocimiento	Conocimiento previo del texto
4	Puntuación en reconocimiento	Nivel de ansiedad en el examen
5	Puntuación en preguntas de verdadero-falso	Conocimiento previo del texto
6	Tipo de ayuda	Conocimiento previo del texto
7	Tipo de ayuda	Nivel de ansiedad en el examen
8	Tipo de relación	Nivel de ansiedad en el examen
9	Tipo de información	Conocimiento previo del texto
10	Tipo de información	Nivel de ansiedad en el examen
3. Escala Estadística		
1	Modelo Completo	Estadístico F
2	Error Completo	Efecto Principal
3	Error Completo	Efecto de Interacción
4	Potencia	Hipótesis Alternativa
5	Potencia	Efecto principal
6	Potencia	Efecto de Interacción
7	Hipótesis Alternativa	Estadístico F
8	Hipótesis Alternativa	Efecto de Interacción
9	Hipótesis Nula	Estadístico F
10	Estadístico F	Comparaciones a posteriori
11	Efecto de Interacción	Tamaño del efecto

Análisis «pathfinder» y multidimensional

A efectos de comparar la bondad de la aplicación anterior, en relación con los procedimientos «tradicionales» experimentales para medir la estructura del conocimiento, procedimos a llevar a cabo un análisis en base al algoritmo «pathfinder», y al escalamiento multidimensional, que procedemos a exponer secuencialmente. La especificación de los resultados de estos dos análisis no es exhaustiva puesto que nuestro objetivo primordial está centrado en el análisis anterior, y éstos se realizan a efectos de comparación global.

Análisis de trayectorias («pathfinder»)

La matriz de proximidad construida en base a las evaluaciones dadas a todos los apareamientos entre conceptos se sometió al algoritmo de «pathfinder» (Schvaneveldt, 1990). Se obtuvieron redes «pathfinder» utilizando valores paramétricos de n-1 (n=número de conceptos) e infinito para q y r. Estos valores generan la red más simple con el menor numero de conexiones posibles.

En cada sesión se obtuvieron redes individuales para el experto y para el examinado al igual que un promedio para cada grupo. El programa utilizado para el escalamiento (Schvaneveldt, 1990) genera diverso tipo de índices tales como el número de nodos y conexiones, la coherencia y una medida de la similaridad entre dos redes. Este índice está determinado por el grado de coincidencia entre las conexiones y se calcula como el número de conexiones comunes dividido por el número de conexiones de ambas menos el número de conexiones comunes. Su valor oscila entre 0 (redes complementarias sin similaridad) y 1 (redes idénticas). En este caso se calculó la similaridad entre cada red individual y la del experto, siendo la principal variable dependiente para la realización de un análisis de varianza entre sujetos realizado en función de la variable independiente «grupo». Debido a que 14 y 22 sujetos de las dos sesiones mostraban índices de similaridad con el experto debidos al azar (p igual o mayor que 0.05) se eliminaron del análisis. A pesar de la selección anterior, el análisis no detectó ningún efecto significativo en ninguna de las dos sesiones, $F(2,60)=1.085$, $p=.344$. $MCe=0.002$, y $F(2,49)=1.432$, $p=.249$. $MCe=0.002$.

Escalamiento multidimensional (INDSCAL)

Los datos individuales pertenecientes a cada una de las condiciones fueron promediados, y la matriz resultante de cada una de las condiciones fue sometida al procedimiento de escalamiento multidimensional de diferencias individuales (INDSCAL, Young y Hamer, 1987) bajo el supuesto de nivel de medición ordinal. INDSCAL asume que se puede extraer una estructura común a partir de una serie de matrices de datos, aunque asume que existen diferencias sistemáticas entre ellas, en este caso porque se han sometido a diferentes condiciones experimentales. El procedimiento fue aplicado independientemente para la prueba 1 y para la prueba 2. Para una solución de dos dimensiones, el algoritmo alcanzó un resultado con un stress respectivo de 0.43 en ambas pruebas. Este stress refleja la dificultad de alcanzar un ajuste mejor.

La interpretación se llevó a cabo de forma dificultosa y por medio de regresión en el que cada dimensión se utilizaba como predictor de cada juicio, independientemente. En la prueba 1 (juicios antes del examen) ambas dimensiones tienen un polo estadístico y otro de tipo psicológico, en relación con el artículo sobre el cual se llevan a cabo los juicios. La dimensión 1 agrupa en su polo más negativo las variables contaminantes mencionadas en el artículo, seguidas por las dependientes. En el otro polo se agrupan los conceptos estadísticos asociados sobre todo con la prueba de efectos (F, alfa por contraste, alfa por experimento, efecto principal, etc.). En la dimensión 2 tenemos en el polo psicológico las variables independientes o hipótesis, y en el otro tenemos conceptos estadísticos asociados con la variabilidad asociada con el modelo restringido (hipótesis nula, error restringido, y modelo restringido). Si fuésemos a denominar de forma única y de forma psicológica a las dos dimensiones a la primera la podríamos denominar factor de «variabilidad medida» (variables dependientes y contaminantes), y al segundo de «variabilidad producida» (variables independientes). En este análisis, y a partir del análisis INDSCAL hay que decir que la importancia de ambas dimensiones es semejante .27 versus .23. El índice de «weirdness» para las distintas condiciones experimentales, es respectivamente 0.04, 0.11, 0.06 (condiciones experimentales), y 0.18 (experto).

Para la prueba 2 las dos dimensiones, aunque se originan a partir de una estructura conceptual común, producen resultados ligeramente distintos (no se olvide que media un examen entre uno y otro). En primer lugar, esta solución tiene pesos más extremos, no siendo las dimensiones tan simétricas como en el caso anterior. La dimensión 1 tiene como un polo la variabilidad medida en su aspecto estadístico (modelo restringido, error restringido) y psicológico (alguna variable contaminante: sexo), mientras que el otro polo tenemos una mezcla de variables dependientes e independientes principalmente. La dimensión 2 agrupa las variables contaminantes (conocimiento previo, ansiedad, sexo) versus conceptos estadísticos asociados con los efectos en el análisis de un experimento. Aquí si tuviésemos que poner nombres tendríamos que llamar a la primera dimensión: dimensión de variabilidad total, y la segunda, sobre todo de variabilidad de error. En este análisis, y a partir de los resultados del INDSCAL hay que decir que la primera dimensión tiene mayor importancia que la segunda por un peso de 0.30 contra 0.23, aunque podemos ver claramente que cada grupo de sujetos (condiciones experimentales distintas) pesan my diferentemente sobre estas dimensiones. Esto se refleja sobre todo en el índice de «weirdness», que es respectivamente 0.01, 0.13, 0.02, y 0.17 para la matriz del experto. Como conclusión hemos de indicar que en general, la aplicación del escalamiento multidimensional fue difícil, el ajuste no muy bueno, siendo la interpretación también difícil.

Discusión

Resumiendo, la aplicación de un procedimiento de estimación de teoría de respuesta al ítem (modelo de crédito parcial generalizado) a unos datos obtenidos en condiciones naturales, que presentaban tres condiciones experimentales distintas de ejercicios prácticos (psicológicos, abstractos, prácticos), muestra que el procedimiento capta cambios específicos en la estimación del conocimiento estructurado (grupo psicológico preexamen) o de forma más generalizada (todos los grupos post-examen, debido al estudio estadístico). Antes de discutir los datos desde el punto de vista psicológico, vamos a analizar comparativamente los distintos procedimientos de medición utilizados.

Se ha podido observar que el método de estimación TRI ha estimado cambios en la habilidad, que ni el escalamiento multidimensional, ni el análisis de trayectorias han detectado. Hay dos razones previsibles para esta diferencia. En primer lugar, el hecho de

que la teoría de respuesta al ítem suministra un procedimiento de medición más preciso de lo que lo hacen cualesquiera de los otros dos procedimientos. No en vano, los modelos TRI se han convertido en los modelos de elección en el desarrollo moderno de tests de aptitudes y rendimiento. Unida a esta razón está el hecho de que la aplicación del modelo TRI permite proceder igual que se construye un test. Esto es, permite un análisis «usual» de ítems. No todos los juicios contribuyen con igual calidad psicométrica a la respuesta final. Sin embargo, tanto en la aplicación del análisis de trayectorias como en el escalamiento multidimensional, tal como son aplicados en este campo de investigación, la única consecuencia es que la solución final pueda generar un residuo mayor o menor, lo que indica un mejor o peor ajuste. En la aplicación del modelo de teoría de respuesta al ítem, se puede eliminar aquellos juicios estructurales que no tienen calidad psicométrica, y garantizar una mejor estimación final. Esta posibilidad, considerada a priori de forma adecuada, permite también resolver uno de los principales inconvenientes de los otros dos métodos; el referido al gran número de apareamientos entre conceptos que hay que llevar a cabo para estimar el nivel de conocimiento de un dominio. En la aplicación aquí hecha, hemos llevado a cabo todos los apareamientos con propósito comparativo. Sin embargo, la aplicación del modelo de crédito parcial generalizado, o de cualquier otro modelo de TRI no requiere la generación de un número de ítems tan grande. Sin embargo, sólo vemos un gran inconveniente en la aplicación de un modelo TRI en el campo de investigación: la garantía del proceso de estimación de parámetros requiere muestras de tamaño a veces considerablemente grande. Este tamaño de la muestra es función lógicamente del número de parámetros a estimar, y de la ocurrencia de patrones distintos de respuesta, entre otros factores. Este hecho puede limitar fuertemente la extensión de estos procedimientos en el campo de la investigación. A pesar de todo, ya ha surgido hasta la fecha una aplicación de alto nivel de un modelo de TRI en el campo de la investigación experimental de memoria (Pirolli y Wilson, 1998).

Desde el punto de vista psicológico, los resultados muestran que antes de que los participantes estudien los conceptos estadísticos, sólo el entrenamiento específico genera una estructura más próxima al experto. Hay que hacer notar que dado que este es un estudio naturalista, tenemos que utilizar los hechos naturales como promotores del cambio conceptual. Durante las sesiones de entrenamiento, los alumnos simplemente atendieron y llevaron a cabo los problemas asignados, pero no llevaron a cabo ningún estudio individual. Este estudio individual sólo se realizó a la hora de preparar el examen, y se reflejó en la segunda prueba. Además, todos ellos fueron presentados de forma expositiva, en una clase teórica lo que impidió aún más la posibilidad de detectar cambios entre las distintas condiciones, salvo por el grupo específicamente sometido a la condición de ejercicios psicológicos. En cambio, los resultados de la segunda prueba muestran que independientemente del entrenamiento específico, los participantes alcanzan un nivel semejante de habilidad; un hecho que podríamos denominar como transferencia por analogía. Ha de hacerse notar que la situación del presente estudio representa la situación ideal para que se de una transferencia de conocimiento por analogía: un conjunto relativamente reducido de conceptos, un contexto restrictivo y con una relación 1:1 bastante clara, y una relación muy directa; es decir, con pocas etapas intermedias, entre el problema inicial (los problemas presentados) y la situación final (el artículo de investigación).

Hay una última pregunta digna de responder que tiene que ver con la posibilidad de transferencia del entrenamiento de un dominio a otro de conocimiento. Antes del examen, sólo el grupo psicológico mostró un conocimiento mayor que los restantes dos grupos, a pesar de que los elementos fundamentales básicos para llevar a cabo los juicios eran los conceptos estadísticos explicados comúnmente. Este dato indicaría una prioridad de lo específico sobre lo general (transferencia del conocimiento estadístico al psicológico). Sin embargo, los datos del segundo test nos invitan a pensar en lo contrario. Dado que este es un estudio hecho en condiciones naturales, hemos de apelar a las «costumbres» del estudiante medio de las facultades de psicología. Puesto que el «estudio» real de un dominio no comienza hasta días antes del examen, es natural pensar que el grado de dominio de los conceptos estadísticos de los dos grupos control es muy débil, y que la familiaridad alcanzada por el grupo experimental a través del entrenamiento específico, es suficiente para suministrarle una ventaja. Y esto es así, puesto que cuando se lleva a cabo el estudio real; es decir, para el examen, los tres grupos quedan igualados. La mejor explicación es aquella que indica que cuando los dominios de conocimiento muestran una relación 1:1 suficientemente clara (diseño-psicológico) la transferencia analógica puede tener lugar fácilmente (Holyoak y Thagard, 1997).

Agradecimientos

La investigación presentada aquí ha sido posible gracias a una ayuda de la Dirección General de Investigación Científica y Técnica (PB/97-1379) del Ministerio de Investigación y Ciencia.

Referencias

Anderson, J. A. (1983). *The architecture of cognition*. Cambridge, MA: Harvard university Press.

Arabie, Ph. (1993): Methodology neither new nor improved. *Contempo - rary Psychology, 38*, 66-67.

Cañas, J. J., Bajo, M. T., y Gonzalvo, P. (1994). Mental models and com - puter programming. *International Journal of Human.Computer Stu - dies, 40*, 795-811.

Collins, A. M., y Loftus, E. F. (1975). A spreading activation of semantic processing. *Psychological Review, 82*, 407-428.

Douglas Carroll, J. (1995). «Minimax length links» of a dissimilarity matrix and minimun spanning trees. *Psychometrika, 60*, 371-374.

Gentner, D. y Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist, 52*, 45-56.

Gillan, D. J., Breedin, y Cooke, N. J. (1992). Network and multidimensional representations of the declarative knowledge of human-computer interface design experts. *International Journal of Man-Machine Stu - dies, 36*, 587-615.

Goldsmith, T. E., Johnson, P. J., y Acton, W. H. (1991). Assessing structural knowledge, *Journal of Educational Psychology, 83*, 88-96.

Gómez, R., Hadfield, O. D., y Housner, L. (1996). Conceptual maps and simulated teaching episodes as indicators of competence in teaching elementary mathematics. *Journal of Educational Psychology, 88*, 572-585.

Gonzalvo, P., Cañas, J. J., y Bajo, M.T. (1994). Structural representations in knowlege acquisition. *Journal of Educational Psychology, 86*, 601-616.

Holyoak, K. J., y Thagard, P. (1997). The analogical mind. *American Psychologist, 52*, 35-44.

Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika, 32, 241-254.

Johnson, P. J., Goldsmith, T. E., y Teague, K. W. (1994). Locus of the predictive advantage in pathfinder-based representations of classroom knowledge. *Journal of Educational Psychology, 86*, 617-626.

Kolodner, J. L. (1997). Educational implications of analogy. *American Psychologist, 52*, 57-66

Maxwell, S. E., y Delaney, H. D. (1990). *Designing experiments and analyzing data. A model comparison perspective*. Belmont, Ca: Wadsworth.

Muraki, E. (1997). A generalizad partial credit model. En W. J. van der Linden, y R. K. Hambleton (1997). *Handbook of Modern Item Response Theory*. New York: Springer Verlag. Pp. 153-168.

Muraki, E., y Bock, R. D. (1997). *PARSCALE. Irt item analysis and test scoring for rating-scale data*. Chicago, Il: SSI Scientific Software.

Novak, J. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching, 27*, 937-949.

Pennington, N. (1987). Stimulus structures and mental representations in expert comprehension of computer programs. *Cognitive Psychology, 19*, 295-341.

Pirolli, P., y Wilson, M. (1998). A theory of the measurement of knowledge content, access, and learning. *Psychological Review, 105*, 58-82.

Reitman, J., y Olson, K. J. (1991). Techniques for representing expert knowledge. En K. A. Ericsson, y J. Smith. *Toward a general theory of expertise. Prospects and Limits*. New York: Cambridge University Press.

Rips, L. J., Shoben, E. J., y Smitt, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12*, 1-20

Rosch (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192-233.

Rowe, A. L., Cooke, N. J., Hall, E. P., y Halgren, T. L. (1996). Toward an on-line knowledge assessment methodology: Building on the relationship between knowing and doing. *Journal of Experimental Psychology: Applied, 2*, 31-47.

Sattath, S. y Tversky, A (1977). Additive similarity trees. *Psykometrika, 42*, 319-345.

Schraagen, J. M. (1993). How experts solve a novel problem in experimental design. *Cognitive Science, 17*, 285.309.

Schoenfeld, A. H. (1989). Teaching mathematical thinking and problem solving. En L. B. Resnick y L. E. Klopfer, eds., *Toward the thinking curriculum: Current cognitive research*. Pp 83-103. Alexandria, VA: Association for the Supervision and Curriculum Development.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization*. Norwood, N. J. : Ablex.

Schvaneveldt, R. W., Cooke, N., Dearholt, D., Durso, F. T., y McDonald, F. T. (1994). A review neither objective nor informative. *Contemporary Psychology, 39*, 100-101.

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427-450.

Wilson, J. M. (1994). Network representations of knowledge about chemical equilibrium: Variations with achievement. *Journal of Research in Science Teaching, 31*, 1133-1147.

Wilson, D. T., Wood, R., y Gibbons, R. (1998). *TESTFACT. Test scoring, item statistics, and item factor analysis*. Chicago, Il: Scientific Software International.

Young, G., y Hamer, R. M. (1987). *Multidimensional scaling: History, theory, and applications*. Hillsdale, N. J.: Erlbaum.

Zwaan, R. A., y Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162-185.

Aceptado el 7 de noviembre de 2000