Tema 5A: Tests adaptativos informatizados. Estructura y desarrollo. Ventajas e inconvenientes.

Licenciatura de Psicología:
Desarrollos actuales de la medición: Aplicaciones en evaluación psicológica.
José Antonio Pérez Gil
Dpto. de Psicología Experimental.
Universidad de Sevilla.

Agradecimientos: a M ªJose Gaviño Alcantarilla, Jorge Perez Valla, Oyvind Ringdal Nygaard y Rocio Romero Blanco.

ÍNDICE

1.- INTRODUCCIÓN

En el contexto de la medición en psicología, el uso del ordenador se hace imprescindible para que los diferentes modelos de la TRI (TEORIA DE RESPUESTA AL ITEM) no queden en en meras expresiones formales.

El apoyo que nos proporciona el ordenador se hace fundamental a la hora de realizar los cálculos matemáticos que son de obligatorio uso para precisar las estimaciones y la combinación de cada ítem a dicha precisión.

Con respecto a la construcción del test los recursos informáticos lo utilizamos para diseñar los ítems a la vez que para crear y mantener bancos de ítems. El ordenador también nos permite seleccionar los ítems adecuados en función de un determinado objetivo o una persona, controlar las veces que se administra cada ítem, determinar si la respuesta es correcta o no, estimar de forma inmediata el rendimiento del sujeto a evaluar y proporcionarle feed-back directo sobre su ejecución después de cada ítem o después del test.

Entre los modelos más importantes de la TRI, cuatro son los más representativos de la aplicación informática:

- Tests convencionales (estos test podrían ser informatizados).
- Elaboración automatizada de informes.
- Construcción automatizada de los tests.
- Test adaptativos informatizados.

Este último es el objeto de nuestro estudio y el que pasamos a desarrollar en este informe.

Un Test adaptativo informatizado consiste en que es un test administrado por ordenador donde la presentación de cada ítem se toma de forma dinámica basándose en la respuesta del sujeto que se examina y en la estimación de su nivel de actividad.

Por otra parte podemos decir que los tests adapatativos son una simbiosis entre los avances informáticos y la TRI. La idea fundamental que los define es que el test a utilizar no es el mismo para todos, sino que se adaptará a su nivel de competencia, a pesar de lo cual el resultado de las mediciones vendrá expresado en la misma métrica, de manera que se ahorra tiempo en la aplicación y se mejora la precisión de medida.

Este tipo de tests presentan tanto ventajas como inconvenientes, además de elaborarse mediante una serie de fases que presentaremos más adelante.

En suma a todo lo comentado anteriormente debemos añadir que los test

adaptativos informatizados necesitan de unos instrumentos para llevarse a cabo , es decir, para su aplicación.

Estos instrumentos son los medios a la hora de planificar, seleccionar y elaborar los test

En cuanto a los criterios de selección del test, tenemos que tener en cuenta, fundamentalmente seis pautas a la hora de seleccionar y aplicar los instrumentos:

1. Establecer cuál es nuestro objetivo y si este se corresponde o no con el objetivo del test. Esto es así porque en la psicología, la medición de rasgo no son observables directamente, por lo que hay que tener muy claro el objetivo de estudio, de esta forma mediremos aquello que nos interese, con lo cual obtendremos validez, no de los principios fundamentales en psicologia.

2. Tener claro la población a la que vamos a aplicar el test y sus características definitorias.

Si optamos por un TAI (TEST ADAPTATIVO INFORMATIZADO) hay otros componentes estresantes que se suman a los típicos y que debemos tener en cuenta los sujetos examinados por una serie de motivos:

a. Al ser una prueba informatizada, el manejo de los ordenadores será variable entre los diferentes sujetos, al igual que su predisposición a usarlo.

b. Los TAIS operativos presentan menos flexibilidad a la hora corregirlos y omitir la respuesta, debido a que es necesaria la confirmación de la mencionada omisión, para pasar el siguiente ítem.

3. Es necesario conocer el lugar donde se llevará a cabo la prueba. Es importante ser conocedor de los recursos con los que contaremos, si la prueba es individual o en grupo. En nuestro caso al elegir un TAI , tendremos que disponer de un lugar con ordenadores que reúnan una serie de características:

- Rapidez de procesamiento y capacidad de razonamiento.
- Buena resolución de pantalla y amplitud de memoria.
- Teclado adecuado.
- Todas las unidades de trabajo serán iguales, tanto en red como en separado.
- Capacidad para poder simplificar en la medida de lo posible la respuesta del sujeto mediante el pecado.
- Si se va a usar el ratón hay que dejar un tiempo de manejo y entrenamiento.

4. Características que ha de tener la persona que elige y administra la prueba. Una persona competente utilizara los tests de forma adecuada, ética y profesional, prestando la debida atención a las necesidades y derechos de los sujetos a evaluar y teniendo en cuenta los motivos por los que utilizamos los tests y el ambiente donde los realizamos.

5. ¿En que medida es un test elegido mejor que otro que mide lo mismo? Hay que distinguir entre TAI y test convencional , para ello propondremos las ventajas e incovenientes de los TAIS, según autores como (Wainer, 1990; Olsen ,1990; Wise y Plake, 1990; Hambleton, Zaal y Pieters, 1990 entre otros):

VENTAJAS:

a. Estandarización; puede crecer en varias estaciones interconectadas como en una red local. Cuando estamos en una situación de examen es una constante donde todas las variables como las imágenes, instrucciones, etc. se producen en las mismas circunstancias.
b. Certificación de títulos y licencias de forma ágil y estandarizadas.
c. Economía por relación en tiempo, nº de ítems y precisión.
d. Cada examinada puede seguir su propio ritmo adecuando la sesión al situación.
e. Posibilidad de actualización y calibración de los ítems renovando aquellos con gran frecuencia de aparición.
f. Diseño a través de programas de dibujo y captura de imágenes mediante escáner aprovechando las cualidades y recursos de la presentación en pantalla. En concreto las animaciones de imágenes, color y el sonido.
g. Feedback automático al conocer al instante los resultados de la prueba.
h. Diagnóstico rápido en el ámbito escolar observando las pautas anómalas de las respuestas.
i. No da lugar a que podamos dar las respuestas por reconocimiento y memoria, ya que no se aplican normalmente los mismos ítems a un examinado que repita la situación.
j. Aunque la duración depende de la longitud de las versiones convencionales las administradas mediante TAI requieren solamente entre 50 % y el 20 % del tiempo original.
k. Han demostrado una gran utilidad para la evaluación de sujetos con problemas físicos y sensoriales donde la fatiga con una larga duración es un fuerte inconveniente.
l. Ausencia de hojas de respuesta lo que facilita el trabajo del examinado focalizando toda su atención en la prueba.
m. Posibilidad de administración convencional tanto de tipo lápiz y papel como informatizada considerando una sola secuencia de ítems.
n. Reducción del número de ítems administrados, puede que sea entre 10 y 20.
o. Simulación de procesos cognitivos y las respuestas esperadas ante un banco

unidimensional.

p. Se puede registrar el tiempo de reacción considerando las limitaciones del ordenador.

INCONVENIENTES:

a. Banco de ítems de gran tamaño periódicamente renovados.
b. La secuencia de dificultad de los ítems no es gradual con lo cual supone que el primero, su presentación, supone un impacto de dificultad.
c. Diferentes pantallas puede dar lugar a deformar los gráficos, imágenes o figuras.
d. Existe poca equivalencia ante pruebas cuya versión convencional es de tiempo limitado.
e. La experiencia nos demuestra la dificultad de la conversión de un TC (TEST CONVENCIONAL) a TA y TAI..
f. La falta de unidimensionalidad en el banco de ítems afecta notablemente al proceso de selección y asignación de puntuación final.
g. Las instrucciones afectan mucho los resultados en este tipo de test.
h. Efectos negativos del teclado convencional a la hora de recoger la respuesta y confirmarla.
i. La computarización supone unos costos adicionales al esfuerzo de diseñar y calibrar unos ítems.
j. No se puede rectificar ni volver a revisar items ya respondidos. Menos flexibilidad que los convencionales, porque lo más normal es que no permitan omitir, diferir o cambiar respuestas.
k. Cualquier ítem anómalo impacta mucho más en el transcurso y los resultados de una sesión de TAI que en los TC.
l. La falta de robustez en la calibración de items producirá malas selecciones de los mismos y la inestabilidad de los parámetros respecto a la muestra.
m. Se exige un alto grado de organización.

6. ¿ Cómo conseguiremos el objetivo?

Si hemos elegido un TAI , hemos de recordar que su algoritmo de trabajo constará de :

Obtener una estimación de la aptitud del sujeto.

Seleccionar de un banco de ítems precalibrados, el ítem que maximice la información, es decir, el que añada más a la función de la información para un nivel de aptitud dado

Administrar y puntuar el ítem, revisar la estimación de la aptitud.

Si la estimación es suficientemente buena, se detiene el proceso. En caso contrario , se vuelve al paso dos.

Ahora comentaremos las fases que se realizan a la hora de construir un test adaptativo informatizado, es decir, para conseguir nuestro objetivo:

2.- FASES PARA LA CONSTRUCCIÓN DE UN TAI

2.1.- FASE I: PLANIFICACIÓN DEL TAI

En esta fase nuestro objetivo es organizar lo que vamos a exponer a lo largo de la construcción , por ello realizamos una serie de pasos. El primero de ellos sería tener claro desde un principio el objetivo final que pretendemos con la evaluación de los sujetos, además de una buena planificación.

Tras ellos debemos elegir el tipo de ítem y las puntuaciones que vamos a utilizar en nuestra evaluación.

Lo más usual es que los ítems utilizados sean Banco de ítems Calibrados (BIC) dicotómicos, unidimesionales y con alternativa múltiple. Este punto también es importante ya que que va a determinar la necesidad del software especializado y la mayor muestra de examinados.

Otros punto importante en la organización, es el tipo de construcción necesario que tendremos que emplear en ella.

Por otro lado tenemos que aceptar una metodología determinada que sea coherente con nuestro proyecto, siempre en función de nuestro objetivo pero respetando la metodología elegida. Por ello debemos mencionar medidas sólidas y de fácil interpretación. Para el examinado una sesión TAI puede darle una impresión equivocada de su nivel, ya que seguramente obtendrá un número semejante d errores y de aciertos.

Por otra parte tenemos que tener en cuenta los costes y beneficios que han implicado, además del equipo informático utilizado frente los actuales. Es cierto que una medida clave de eficiencia converge en un ahorro de un 50 % de tiempo aproximado por examinado a favor de los TAIs frente a los tests de lápiz y papel.

Lo más habitual es efectuar el DAP (Diseño de Anclaje de pruebas) con pruebas de tipo test de láminas de lápiz y papel basadas en ítems de alternativas múltiples puntuados dicotómicamente.

Por otro lado muchos tipos de ítems parecen incompatibles con los TAIS por la dificultad de corrección e implementación en un software. Sin embargo, existen soluciones alternativas, como la modalidad test adaptativo parcialmente informatizado compatible con esta clase de item y entornos variables de examen mediante ordenador portátil.

Una decisión que determinará el resto del proceso sería la adecuación del contenido a las condiciones de la TRI . (Teoría de la respuesta al Item).

Fabricar un BIC unidimensional satisface una serie de requerimiento de la TRI

como puede ser la ausencia de anomalías producidas por otras características de los examinados (funcionamiento diferencial de los ítems), ajuste de respuesta de cada ítem a un modelo, etc.., pero sin embargo la elección de un contenido ajustado a la TRI facilita este objetivo.

Las complicaciones aumentan en el caso de considerar la velocidad como un factor clave en la puntuación del examinado.

2.2.- FASE II : PRODUCCIÓN DEL BANCO DE ÍTEMS.

En esta parte del procedimiento, el principal objetivo es crear una amplia colección de reactivos de forma estandarizada ,con una dificultad mayor al los de TC, y que el contenido se forme a partir de las reglas de generación de ítems.

Respecto al contenido, las reglas de generación de los ítems pueden basarse en el análisis lógico de ítems mediante diagramas de flujo, grafos o algoritmos que puedan generar de modo sistemático reactivos completos. Esto lo que hace es reducir los riesgos de anomalías y puede facilitar el mantenimiento del futuro BIC.

Estas reglas además de repercutir en la dimensionalidad y validez del contenido del BIC, también permiten identificar a los descriptores más útiles de los ítems.

Podemos tener dos caminos, que el objetivo sea de medición de aptitudes o que sea de evaluación del conocimiento y rendimiento. En caso de que sea de medición de aptitudes será bastante importante y necesario disponer de un modelo cognitivo de referencia, para guiarnos y así poder hacer mejor nuestro trabajo, lo que contribuirá a la creación de nuevos ítems y de reglas de generación. Por otra parte si lo que nos interesa es la evaluación del conocimiento y rendimiento, será muy adecuado diseñar una tabla de especificaciones que pueda combinar las principales áreas de contenido y taxonomía de conocimiento que mejor responda a los objetivos de la evaluación.

Por otro lado podemos señalar que la disponibilidad de un peritaje de expertos sobre la dificultad estimada y descripción objetiva de características del reactivo es una pieza clave para abordar estas posibilidades.

La unión de todos los descriptores codificados numéricamente bajo criterios comunes producirá un vector descriptor del item que durante el resto del proceso pasará a identificar y caracterizar cada pieza del BI.(BANCO DE ITEMS)

Señalaremos también que un mismo nivel de complejidad teórica puede obtenerse a partir de diferentes VDI (variables independientes).

En cuanto a la determinación del tamaño del banco de ítems, sabemos que lo deseable sería administrar a todos los examinados el BI completo,lo que ocurre que se corre el riesgo de potenciar efectos negativos que deteriorarán la calidad de las respuestas, como puede ser la fatiga, resonancias, fatiga, desmotivación, etc…Esta motivación no solo afecta

a bancos grandes sino también a los bancos pequeños que exigen más esfuerzo o más tiempo de respuesta, y un TAI es eficiente en la medida en que disponga de muchos ítems donde escoger, ya que se acentuarán los factores no controlados que puedan afectar a la estimación de la capacidad.

Hay que organizar un procedimiento en el cual se distribuya los ítems de la respuesta en diversos bloques para obtener dichas respuestas.

Un bloque no es una prueba del DAP, sino el conjunto de respuestas de los examinados que lo realizan. Cada bloque se administra convencionalmente.con tests de lápiz y papel o por pantalla, bajo unas mismas condiciones, y sin límite temporal, se les reparte a uno o más examinados, de tal manera que se recojan suficientes pautas de respuesta con las que analizar, calibrar y equiparar los ítems. Este procedimiento es lo que constituye el DAP, y vendrá marcado por factores como puede ser la cantidad , disponibilidad de los examinados, tipo de equiparación aplicado, tiempo de administración, fatiga de examen, etc…

El DAP más adecuado es dependiendo del objetivo y del perfil deseado de la FI (función de información) del banco.Si hay pocos ítems en cada bloque se necesitarán más muestras y poco tiempo en cada una.En cuanto a los ítems escogidos, son los que en ellos se depositan más confianza que ninguno.Para que efectuar equiparación con garantías tiene que haber como mínino dos ítems hasta superar el 50% del bloque. La distribución de dificultades que existe en cada bloque, depende de si la dificultad de los bloques es diferente o bien en todos se mantiene una representación de ítems con dificultad similar, el DAP determinara una equiparación vertical u horizontal .

Los DAP pueden variar notablemente desde estructuras muy simples hasta diseños con bloques encadenados y en red mucho más complejos. Éstos se clasifican en dos grandes modalidades dependiendo si se mantienen unos mismos ítems ancla en todos los bloques o en cambio van cambiando según los bloques que unen.

Por otro lado, podemos comentar que el tamaño de la muestra dependerá en parte del modelo TRI, del DAP y de la infraestructura disponibles. Por lo general el modelo de 1PM precisa menos examinados, mientras que en el de 3PM sucede todo lo contrario. La muestra debe reunir individuos que se distribuyan a lo largo de todos los niveles de la futura escala de capacidad, lo que afectará a la motivación e implicaciones del resultado del examen.

2.3.- FASE III: CALIBRACIÓN Y ENSAMBLADO DEL BANCO DE ITEMS

Obtención de datos y análisis previos por bloques

Lo primero que hemos de hacer en esta fase y, antes de calibrar los ítems frente al gran volumen de información manejado en un DAP, es efectuar una auditoria de las respuestas de los sujetos. Así aliviaremos los siguientes trabajos de calibración equiparación y análisis sobre sesgo de los ítems. También es conveniente que verifiquemos cualquier anomalía, tanto en ítems como en los examinados y depurar filas y columnas sospechosas de cada matriz de datos o bloque del DAP. Pasamos a describir este control que presenta tres frentes de acción:

a) El primero de ellos consistiría en filtrar la captura-obtención de datos a fin de evitar la tabulación de protocolos anómalos de examinados. Para ello llevaríamos a cabo una serie de pasos:

- Verificar efectos tipos Garbage In-Garbage (GIGO) (Doménech, Losilla y Portell, 1997), para identificar y evitar toda la serie de amenazas y problemas en la calidad de los datos, producidos durante el registro y gestión de las matrices de cada prueba (errores de tabulación por fatiga, efectos del escáner o daños durante la transferencia y manejo de ficheros…).

- Registrar las pautas de respuesta en bruto, sin correcciones, incluyendo dobles marcas, omisiones y reactivos no abordados por falta de tiempo. Más adelante esto permitirá detectar los ítems defectuosos (flawed ítems).

- Convertir los datos brutos en corregidos y posteriormente realizar un análisis exploratorio de los datos para conocer la forma de las distribuciones de las puntuaciones.

b) El segundo de los frentes de acción que debemos llevar a cabo consiste en realizar un serie de análisis convencionales de cada prueba que permitan localizar ítems incompatibles con los modelos TRI. Los resultados serán especialmente importantes para los ítems de anclaje, verificando los siguientes aspectos:

- Distribución de frecuencia de elección de los distractores que pueda indicar un mal diseño, confusión, resonancias con el enunciado (distractores regulados) o una probabilidad inesperada de acierto por conjetura. Esto repercutirá sensiblemente en los cálculos necesarios para la detección de individuos sospechosos de copia o entrenamiento (Belleza y Belleza 1989).

- Si es factible, realizaremos un análisis de frecuencia y distribución de las categorías de respuestas erróneas en ítems abiertos.

- Después realizaremos una discriminación convencional del ítem.

- Estableceremos una posible detección de ítems defectuosos con ninguna o varias respuestas correctas (Potenza y Stocking, 1997).

- No solo es importante analizar los aciertos y errores sino también el comportamiento en cada ítem de omisión en relación a la capacidad de los examinados. Para ello es conveniente hacer una inspección del perfil de acierto y perfil de omisión de ítem (Renom, 1995). Más adelante los perfiles de omisión a los ítems se valoraran junto a las CC (curvas características) de los ítems de la omisión a los ítems y gráficos de residuales estandardizados (IDEA I+D, 1992) para localizar rangos de θ conflictivos. Este análisis es extensivo al perfil de doble marca y en especial a cada uno de los distractores (Wainer, 1989). Para una buena equiparación es deseable que la forma del perfil de omisión del ítem, el perfil de doble marca y los perfiles de los distractores sean parecidos para unos mismos ítems ancla en los diversos bloques donde se encuentren.

- Por último dejar claro que en pruebas de rendimiento es importante que relacionemos cualquier deficiencia de los ítems con su VDI. En el caso de medir otras aptitudes esta conexión afectara a los componentes implicados del modelo cognitivo.

c) En tercer lugar y ya para finalizar debemos de hacer una verificación de las pautas de respuesta de los examinados y tratar de detectar los siguientes aspectos:

- PAR(valoración del parámetro de la persona) sistemáticos (Mejler y Sijtma, 1995) que nos lleven a localizar individuos con respuestas anómalas (aciertan lo difícil y fallan lo fácil, incoherencia…). Esta exploración debe ser previa y además necesaria para la efectuada desde TRI mediante el ajuste con la curva esperada de respuesta de persona (CERP) (Trabin y Weis, 1983; Nering y Meijer, 1998).

- Parejas de examinados con un patrón de errores muy similar, especialmente en los ítems difíciles. En función de su proximidad física (coordenadas fila-columna) durante el examen (Scrutiny!, ASC, 1997) y las repercusiones del resultado, puede tratarse de individuos que poseen un conocimiento similar sobre el contenido de los ítems o bien que adoptan una misma estrategia de respuesta (entrenamiento) o, en última instancia, de sospechosos de copia.

- Al igual que antes con los ítems, el resultado de estas comprobaciones debe relacionarse con la información personal disponible de los examinados y la submuestra del DAP a la que pertenecen.

Las tareas b y c hemos de combinarlas y ejecutarlas cíclicamente, ya que la interacción ítems-sujetos hará que cualquier depuración afecte al conjunto de datos. Sólo así podremos afrontar la calibración de ítems con una cierta seguridad.

Calibración, ajuste y valoración de la dimensionalidad

Una vez que hayamos revisado y depurado las matrices de datos es necesario proceder a calibrar los ítems de cada bloque valorando su dimensionalidad y ajuste con el modelo escogido o que mejor funcione. De nuevo aquí será importante observar especialmente el comportamiento de los ítems ancla.

Para la calibración de ítems existen diferentes procedimientos de estimación de parámetros y software que los implementa (Baker, 1992; López Pina, 1995). El objetivo en todos será obtener unos valores invariantes y robustos ajustados a la CC del ítem obtenida según el modelo TRI escogido (López Pina, 1996). Tanto o más importante será verificar la unidimensionalidad de los reactivos.

Se han propuesto muchos métodos para evaluar la presencia de un solo rasgo influyente en la respuesta a los ítems. Hattle (1985) los agrupa respecto a si el concepto se apoya en los patrones de respuesta, la fiabilidad, en el análisis factorial, en el análisis en componentes principales o en modelos de la TRI.

Si las respuestas a los ítems sean intuidas básicamente por un solo factor, entonces las pautas de respuesta a un conjunto de ítems deben ser coherentes con las características de éstos respecto a dicho factor. El caso más evidente, pero también más difícil de obtener, es el ajuste del patrón de respuesta a un escala upo Guttman, suponiendo la presencia de un único factor que varía en dificultad. En este sentido, la presencia de PAR puede invalidar el supuesta unidimensionalidad (Reise y Waller, 1993), por lo que puede resultar interesante, si no necesario, un estudio previo de este tipo de desajuste antes de evaluar la unidimensionalidad con alguno de los otros medios disponibles.

En el marco de la TRI también puede plantearse la detección de PAR (Meijler y Sijtma, 1995). De hecho, desde esta perspectiva la valoración de unidimensionalidad en el BI pasa por obtener un buen ajuste del modelo unidimensional elegido a los datos, y éste no es posible si abundan las respuestas anómalas. La facilidad práctica de esta estrategia la convierte en una de las más utilidades en la evaluación de la unidimensionalidad (Cuesta, 1996).

Por otro lado, es fácil suponer que si en un conjunto de ítems existen unas relaciones consistentes, ese conjunto será homogéneo, y por lo tanto unidimensional. Este razonamiento ha dado pie a la utilización de los índices de consistencia interna, sobre todo el alfa de Cronbach, como detectores de la unidimensionalidad. Curiosamente, si bien es cierto que cuando no hay mucha consistencia interna no puede haber un solo factor subyacente al conjunto de ítems, también lo es que un grupo de ítems multidimensionales también puede ser consistente, por lo que esta forma de valorar la unidimensionalidad presenta grandes limitaciones.

La evaluación de unidimensionalidad desde el modelo factorial, incluyendo en él tanto el análisis factorial como el análisis de componentes principales, consiste en determinar hasta qué punto una estructura de covariación entre ítems puede resumirse en un único factor. Está claro que la representatividad máxima ocurriría cuando el primer factor

explicase el 100% de la varianza, pero desgraciadamente esta no es una situación realista, de manera que atendiendo a resultados empíricos, algunos autores ya consideran aceptables porcentajes del orden 20 o el 40% de la varianza. De cualquier modo, es importante valorar la evolución de la varianza explicada por los distintos factores. A modo de ejemplo, señalar que la confianza en la unidimensionalidad es mayor cuanto mayor sea la diferencia de varianza explicada por los factores primero y segundo.

Hemos de tener en cuenta que en los modelos de TRI unidimensionales más utilizados el tipo de datos que se analiza es de carácter dicotómico, y en este caso resulta más adecuado analizar la matriz de correlaciones tetracóricas entre los ítems. (Ferrando, 1996), describe el cálculo y la utilidad de algunos índices que pueden utilizarse en estos casos. Una extensión de la dimensionalidad del BI se encuentra cuando se compara las ejecuciones de grupos diferentes ante el mismo BI. En este caso, sería indicadora de multidimensionalidad la presencia, en grupos distintos, de un funcionamiento diferencial en los ítems (DIF). Los grupos se definen en función de variables que a priori puedan ser influyentes en las respuestas a los ítems. En el caso de banco de sumas, por ejemplo, podría tener sentido la evaluación del DIF respecto a grupos con niveles de formación escolar diferentes.

Si como resultado del análisis de la dimensionalidad y del DIF concluimos que el banco es heterogéneo o multidimensional, se suele considerar adecuado que, antes de la calibración definitiva de los ítems, se eliminen del banco aquellos que rompen con las condiciones deseadas.

Disponiendo de suficientes muestra, una práctica recomendable es calibrar dos veces los ítems del bloque, empleando en cada una la mitad de examinados. Luego se comparan, según el modelo, las respectivas estimaciones de a, b y c en busca de linealidad. Otra versión alternativa consiste en efectuar, y comparar, dobles estimaciones de θ para todos los examinados del bloque empleando la mitad de ítems en la obtención de cada una.

Proceso de equiparación de parámetros

El tamaño del BI necesario para construir un TAI hace imposible la administración a un solo grupo de todos los ítems del banco. Afortunadamente, pueden plantearse estrategias con el fin obtener ítems calibrados en la misma métrica, sin necesidad de administrar todos ellos a los mismas personas, dichas estrategias incluyen una minuciosa planificación que comienza con el DAP y termina con la equiparación de los parámetros de los ítems mediante alguno de los mucho métodos disponibles.

Aunque es posible realizar la equiparación de los parámetros de diferentes ítems contestados por distintas muestras de sujetos que se consideren representativas de la misma población, lo más habitual es prever la posibilidad de que algunos ítems sean contestados por todos los sujetos, que algunos sujetos contesten a todos los ítems o una combinación de ambas estrategias. El objetivo, en cualquier caso, es tener alguna referencia común que sirva de anclaje en la equivalencia de las distintas métricas. En un DAP con diseño de anclaje de ítems, la práctica habitual es que los ítems ancla son los mismos en todas las formas del test.

En un DAP para un banco de sumas con formas de dificultad comparable, el proceso de equiparación se hace más sencillo y se le denomina horizontal. Cuando la equiparación sea con formas de dificultad distintas, la equiparación se denomina vertical.

Respecto a los métodos de equiparación, podemos distinguir los métodos clásicos, que incluyen la equiparación lineal y el equipercentil, y los métodos basados en los momentos o en la curva característica de la TRI, como los métodos basados en los momentos o en la curva característica del test. Queremos destacar aquí, por su simplicidad y resultados comparables a otros métodos (Navas, 1996), el de la calibración concurrente (Hamblenton y Swaminahan, 1985), que simplemente consiste en calibras simultáneamente todo el conjunto de ítems a partir de la matriz de datos obtenida mediante un diseño de anclaje de ítems, tras suponer que los sujetos no alcanzaros a responder a los ítems que no fueron administrados.

2.4.- FASE 4: IMPLEMENTACIÓN Y EJECUCIÓN DEL TEST ADAPTATIVO INFORMATIZADO

Implementación del banco de items en un software TAI

Una vez que sabemos las posibilidades del banco de items, entramos en otra fase cuyo objetivo en principio es implementar los items, con sus características y parámetros, dentro de un software de administración determinado (Wainer, 1990). Este primer objetivo puede hacerse más fácilmente dependiendo de si la administración convencional de los bloques del DAP fue realizadas con test de lápiz y papel o a través del ordenador.

En los últimos diez años el software más usado es el denominado MicroCAT de Assessment System Corp. Esta herramienta de trabajo nos permite la calibración de items desde los principales modelos de la TRI, además de administrarlos en función de diferentes algoritmos adaptativos. La estructuración en módulos consigue cubrir las principales tareas que forman un proyecto TAI.

Otros programas que nos resultan más familiares, como son los del entorno de Windows (Fast Test Pro o LearnStar), nos permiten la administración de un TAI a través de items multimedia. Incluso uno de ellos (LearnStar) puede hacer que las administraciones se realicen vía internet, aumentando la expansión y potencialidad de los TAIs.

Aplicado a nuestro ámbito universitario se han desarrollado dos aplicaciones a los TAI: una sería el DEMOTAC (programa informatico aplicado para realizar la rutina de un BI) (Martínez y Renom, 1993) y ADTEST (A computer-adaptive test) (Revuelta, Ponsoda y Olea, 1993), ambas con fines didácticos y de investigación.

Podrían darse una serie de consejos a efectos operativos e independientes de cada software a la hora de administrar los items a través del ordenador:

- mostrar en cada pantalla el tiempo límite del que se dispone para

contestar cada uno de los items, si es que hay tal límite y exponerlo además en minutos
- señalar la cantidad de items que se han contestado
- establercer entre la presentación de un item y el que le sigue, una separación mínima de 0,5 segundos
- el espacio que ocupa un item debe ser de una pantalla como máximo, sin que se tengan que usar barras de desplazamiento
- por último simplificar el teclado para que sólo esten presentes las teclas necesarias e imprescindibles para cumplimentar el TAI.

Selección del algoritmo adaptativo

Otra parte fundamental del mecanismo de un TAI sería la correspondiente a la elección del algoritmo, en la que: (1) inico, (2) continuación, (3) final.

Hay muchas maneras de poder combinar estos tres momentos, y cada una de estas combinaciones significará una u otra forma de establecer la evaluación. A lo largo de la historia se ha ido pasando de mecanismos rígidos a otros mucho más flexibles, por lo que se dice que las estrategias adaptativas han evolucionado. Hay tres procedimientos característicos a la hora de realizar una estimación:

1.- Binivel: se le pasa al sujeto una serie de items empezando por una fácil y acabando por uno de mayor dificultad. Esta serie de items formaran lo que se denomina primer nivel, y deben dilucidar la zona delimitada entre el último acierto y el primer error. Después se pasa otra serie de items que corresponden a esa zona delimitada para establecer así unos items de nivel medio.

2.- Piramidal y ramificado: lo primero que se hace es administrar un item de dificultad media, después dependiendo de si se ha acertado o no, se pasa otro item siguiendo la estructura rígida del árbol.

Es necesario dejar claro que estos dos procedimientos se consideran demasiado rígidos por lo que se busca otro procedimiento que resuelva este problema. Además cuentan con el inconveniente de la excesiva derivación que se le da tanto a los items de nivel 1 como a los de las primeras ramificaciones, ya que al divulgarse sus resulatados pueden dar a conocer sus respuetas acertadas a los nuevos sujetos que vayan a examinarse.

3.- Ramificación variable: aunque la estimación a la que lleguemos con este proceso sea muy parecida a la de los dos anteriores, el camino seguido para ello es menos determinístico y tolera mejor el hecho de que haya aciertos o errores que no sean consistentes.

Hablando ya en términos generales, los algoritmos adaptativos no dependen exclusivamente de la forma de seleccionar el siguiente item. Cada momento tiene sus variantes:

INICIO: la manera en que se comience un test influye en el primer item, pero

además en las instrucciones y ejemplos que se exponen antes de comenzar el test. El primer item elegido puede ser de una u otra forma en función del procedimiento a seguir posteriormente. Un item de dificultad media si es piramidal, uno sencillo si es binivel o uno al azar si es flexible.
CONTINUACIÓN: aquí se aprecia si es o no posible omitir y rectificar las respuestas. Si se pueden realizar omisiones, podrá optarse por ignorarlas, penalizarlas o considerarlas como un error más. El dejar omitir conlleva riesgos como el hecho de que el sujeto espere a adquirir confianza antes de empezar a responder o que se produzca una rápida divulgación del BIC. Por otro lado puede permitirse la rectificación de forma dinámica, bien durante la sesión o después de acabar el test.

Por norma general los TAIs no permiten ni omitir ni rectificar, y esta se considera una de las causas para que estos tipos de test no terminen de aceptarse publicamente.

Otro de los aspectos a decidir es el tiempo máximo que estará un item y el control de exposición de este. En los TAIs suele ponerse límite de respuesta para cada item. En lo referente al control de exposición, lo que se hace es condicionar el método de selección del siguiente item, según su frecuencia de aparición acumulada.

También es necesario decidir si se informará a los sujetos del resultado obtenido en la prueba o no. Esta decisión influirá notoriamente en el itinerario de presentación.

FINAL: puede haber diversos motivos por los que se termina un TAI, los cuales dan lugar a reflexionar.
Después de un número de presentaciones: se verán aumentadas las diferencias en el error típico de medida (ETM) de los sujetos.
Tras agotar un tiempo límite impuesto: la puntuación del sujeto dependerá de las diferencias individuales debidas al tiempo de respuesta.
Por detección de PAR: un TAI bien hecho es un buen detector de PAR. El ETM tenderá a crecer o aumentar si se produce un itinerario anómalo de respuestas. Pude que se combinen algunos de estos criterios y que por ello tenga que finalizar el TAI.

2.5.- FASE 5: EXPLOTACIÓN Y GESTIÓN DEL TAI

La sesión de evaluación

Después de haber metido el contenido del banco de items en uno de los softwares mencionados anteriormente y haber elegido el procedimiento de administración de la prueba que consideremos más adecuado, nos queda por último conseguir un ambiente de tranquilidad en el espacio en el que se ejecutará la prueba. Es importante para el buen desarrollo del examen, que el sujeto conozca las instrucciones y este familiarizado con el teclado y la mecánica de la prueba (Sands, Walter y McBride, 1997; Wainer, 1990). Para

ello podemos pasar primero unos items de prueba para ayudarles a la toma de contacto. La meta de las instrucciones es que sepan el tiempo que disponen para contestar cada item, si pueden o no omitir, si puede haber más de una o ninguna respuesta correcta y cuanto durará la sesión en total.

Mientras realiza la prueba del TAI el sujeto se irá encontrando con items de diversa dificultad que se ajustarán a su nivel, por lo que encontrará un dominio en torno al 50%. Esto paso no se obtenía anteriormente cuando nos disponíamos a calibrar los items, por lo que apreciamos una importante diferencia. También aclarar que el resultado que se obtiene en un TAI y lo que el sujeto cree que ha obtenido no tienen porque coincidir siempre.

Con respecto a los productos que constituyen una sesión TAI nos encontramos:

- se trata de un patrón de respuestas que no se debe perder porque tiene información acerca del itinerario que determinará posteriormente PAR o bien mantener y renovar el banco de items
- un perfil de ETM que esperamos que vaya en descenso a la vez que se van dando las respuestas consecutivas.
- el número de items administrados
- el tiempo que se emplea en realizar cada item y el total de la sesión una capacidad que se estima y su transformación en puntuaciones comprensibles para todo el mundo.

Según Nering (1997), cuando nos disponemos a detectar un PAR, si utilizamos los índices estandarizados de ajuste de individuos obtenidos mediante el TAI, puede haber problemas debido a que estos índices no se distribuyen normalmente en este contexto específico. La detección de estas anomalías ha supuesto una línea de investigación en los últimos años.

Fiabilidad y validación de las puntuaciones

Aquello que en los TC correspondía al cálculo de la fiabilidad de las puntuaciones por medio de coeficientes de equivalencia, estabilidad o consistencia, es lo que los TAIs han centrado sobre los ETM. Este ETM oscila en el continuo en función de la cantidad y calidad de los items, por ello tenemos la posibilidad de efectuarles estimaciones con diferentes precisiones acerca de una determinada capacidad que posean los sujetos. Este ejemplo deja patente diferencias entre el TAI y el TC: si a un grupo de sujetos se le pasan dos test que consideramos paralelos, obtendremos un coeficiente de fiabilidad que nos parecerá bajo debido a que las estimaciones de una capacidad suelen ser poco precisas. Si por el contrario se usa el mismo banco de items pero las estimaciones de la capacidad son precisas el coeficiente obtenido será más alto.

Pasando ya a la validación de las puntuaciones, permanece la estructura clásica de tres niveles.

1.- Validez de contenido: implica que se haga un esfuerzo continuado, aunque se encuentra muy asociado al proceso de construcción del BIC.

2.- Validez de criterio: se realiza en dos niveles. En el primero de ellos se compara lo obtenido en el TAI con lo que se obtuvo en la versión que se pretende cambiar. En el segundo se compara el TAI con la predicción de factores y variables de éxito (Sands, Walter y McBride, 1997) relacionadas con los objetivos de selección, certificación y ahorro que muestren mejoras sobre los TC.

3.- Validez de constructo: incorpora todos aquellos factores que puedan incidir en la obtención de las puntuaciones, incluyendo tanto el marco teórico, como los efectos del teclado, estrategias adaptativas usadas...

Para demostrar sus ventajas sobre los test clásicos los TAIs han tenido que justificarse usando conceptos de la psicometría tradicional como fiabilidad y validez.

2.6.- FASE 6: MANTENIMIENTO Y RENOVACIÓN DEL TAI

El mantenimiento tanto del TAI como del banco de items comienza en su misma explotación (Mills y Stocking, 1996; Stocking, 1996; Stocking, 1997; Umar, 1997; Wise, Curran y McBride, 1997). Muchos proyectos terminan en la fase 5. En cambio otros pasan a esta fase por dos motivos: actualizar loa parámetros de los items y renovar los más desgastados por otros que resulten similares, pudiendo ser la renovación total o parcial dependiendo de si se reemplaza el item totalmente o si parte de su contenido se recicla. Otra cosa sería detectar un item defectuoso.

Para al mantenimiento y renovación de los BIC o los TAIs hay que tener en cuenta:

- prever un historial de cada item (Gershon, 1990) teniendo en cuenta por ejemplo el tiempo medio de exposición. Si el tiempo de latencia es muy pequeño o demasiado largo puede significar que hay alguna anomalía.
  los items que más se exponen suelen ser los mejores, por lo que sustituirlos supondrá una esfuerzo mayor aún que la producción del banco de items. También puede deteriorarse la explotación del TAI (Mills y Stoicking, 1996).
- la calibración puede verse dificultada debido a la no coincidencia de una sesión TAI con el banco de items. Los problemas de mantenimiento más comunes son: dificultad para aprovechar la información obtenida tras diversas sesiones y los problemas de calibración de los items implantados de forma dinámica en el banco de items.
- si se quiere renovar será necesario crear nuevos items para que aparezcan durante la sesión y que además no intervengan en el resultado. La meta por tanto es conseguir una matriz de datos sólida para realizar la calibración aplicando un sistema de anclaje-equiparación dinámicos. El problema de la calibración aquí es el sesgo que supone es estar determinada por la distribución de puntuaciones.
- detectar itinerarios frecuentes, sean o no esperados, puede modificar la valoración de exposición de un item, además de la conveniencia de emplear las pautas de respuestas de los sujetos en la nueva calibración
  la base de la renovación se encuentra en analizar los items que sean más y

menos informativos para así poder aproximar el valor de los parámetros del item y precisar mejor su lugar de inserción y presentación en el banco.

Para finalizar decir que lo ideal sería poseer un generador de items más que un banco (Millman, 1984).

Valor añadido de un proyecto TAI

Hay otras ventajas que proporcionan los TAI aparte del ahorro de tiempo y de items. Poner en funcionamiento un TAI resulta complicado, tanto por la variedad de elementos a trabajar como por el nivel de profundidad de las teorías en que se basa. Los TAIs permiten conocer mejor las estrategias y mecanismos de las respuestas del examinado que los TC debido a que crean una mayor flexibilidad. Un TAI producirá un nivel de estandarización en las medidas determinadas mucho mayor que el que produciría un TC.

Una vez explicado con detalle cómo se construye un TAI, pasamos a señalar una serie de investigaciones sobre el tema; posteriormente, y ya para terminar, hablaremos de los test on-line y para ello resumiremos un estudio que nos cuenta la situación de este tipo de aplicación.

3.- INVESTIGACIONES EN TEST ADAPTATIVOS INFORMATIZADOS

El siguiente apartado pretende informar sobre los problemas más importantes que se han detectado en su utilización, así como las líneas de investigación que se siguen en la actualidad para buscar las mejores soluciones.

Algunas tiene que ver con la mejora de las condiciones de aplicación, algo que resulta imprescindible si queremos dar el salto desde la investigación hasta la implicación generalizada de TAIS en diversos contextos de evaluación. También tienen mucho que ver con esto determinadas modificaciones en los algoritmo de selección de ítems, que se realizan para adecuarse mejor a las necesidades deíndole aplicada que se demandan en situaciones reales de evaluación o los intentos por incorporar nuevos modelos psicométricos y diseñar TAIS para nuevos contenido.

Lo que resulta cada vez más evidente es la utilización de estrategias de simulación como metodología fundamental de investigación sobre las propiedades métricas de los TAIS, complementaria a la índole experimental y en algunos casos imprescindible.

Entre los múltiples núcleos de investigación a tratar, consideramos que son dos los más relevantes (el tema del control de la exposición y los procedimientos para establecer restricciones en la selección). La relevancia la valoramos desde un punto de vista aplicado ya que es difícil imaginar un TAI aplicado en un contexto real que no incorpore en el algoritmo restricciones sobre ambos temas, e investigador.

En cuanto a la metodología de la simulación es importante hablar ya que los trabajos de investigación sobre los TAI pueden realizarse con sujetos reales o simulados.

En estos últimos es donde la aplicación de los test se realiza únicamente con el ordenador, simulando las respuestas que darían los sujetos a los items. Cabe destacar al menos dos ventajas importante de la simulación: la mayor rapidez y economía con que pueden realizarse las investigaciones y la posibilidad de conocer el verdadero nivel de habilidad de los sujetos evaluados. La metodología de simulación se utiliza ampliamente en estadísticas para estudiar aquellos sistemas que, por su complejidad, no tienen soluciones analíticas.

Para simular la respuesta de un sujeto el primer paso consiste en terminar el verdadero nivel de habilidad. Este valor puede fijarse por el experimentador o tomarse aleatoriamente de una determinada distribución como la norma estandarizada. A continuación se selecciona el item que se va a administrar y se calcula la probabilidad de cada una de las posibles respuestas, utilizando el valor verdadero de habilidad. El tercer paso consiste en obtener un valor aleatorio uniforme entre 0 y 1, y determinar la respuesta al item comparando este valor con las probabilidades anteriores.

Después de aplicar un número determinado de items según el procedimiento establecido en el algoritmo, puede estimarse el nivel de habilidad del sujeto simulado, comparar las estimaciones con los parámetros preestablecidos o estudiar la eficiencia del TAI.

Los TAI se basan en el principio de adecuar los más posible el contenido de los tets a las características de cada sujeto evaluado. A partir de un amplio campo de item se seleccionan aquellos que son más apropiados a cada sujeto básicamente aquellos cuyos nivel de dificultad se aproxima al nivel de habilidad del sujeto. Sin embargo en la práctica es frecuente encontrar que hay items que aparecen en una gran cantidad de tets, mientras que en otros apenas son utilizados por ello es necesario el control de la exposición de los items. La frecuencia con que se utiliza un items en distintas aplicaciones de un tets se denomina *tasa de exposición*, y viene determinada en gran medida por las propiedades psicométricas del banco de items.

En la literatura aparecen descritos distintos métodos de control de la exposición, en los que se persigue dos objetivos fundamentales:

1. Prevenir la sobreexposición de algunos items, evitar que se usen en demasiados tets lo que puede representar una amenaza para la validez de la prueba.
2. Incrementar la tasa de exposición de los items menos utilizados. Con estos se consigue aumentar la variedad de items que se administran en diferentes tets, lo que resulta deseable conociendo los costes de todo tipo que supone el desarrollo de bancos de items.

Cualquier otro método de control de la exposición implica que no se selecciona en todo momento el items más adecuado (más informativo) para cada sujeto. Por lo tanto, cabe

esperar que en los TAIS que incorporan un método de control se produzca cierto grado de imprecisión. En los métodos de control de las tasas de exposición puede clasificarse en dos grandes grupos:

1 Métodos que añaden un componente aleatorio al método de máxima información: por ejemplo, uno consistiría en seleccionar el primer items del test al azar entre los cincos más informativos, el segundo entre los cuatro, el tercero entre los tres, el cuarto entre los dos más informativos y a partir del quinto items se administran el más informativo

2 Método de control directo de la tasa de exposición: la segunda categoría de métodos son aquellos que asignan parámetros a cada uno de los items para controlar de forma directa su tasa de exposición. El método original dentro de éste grupo fue propuesta por Sympson y Hetter (1985). El método de Sympson y Hetter, y sus derivados tienen la ventaja de que permiten un control directo de la tasa máxima de exposición y la posibilidad de adaptarlo para recoger todas las posibles características del tets. De este modo es posible fijar el valor máximo de la tasa en el total del tets, controlando por habilidad y teniendo en cuenta las distintas restricciones no psicométricas en las selecciones de items. Algunas dificultades a la hora de la aplicación practica son: la complejidad de los métodos y del proceso de asignación de los parámetros de exposición, especialmente para las extensiones del método original.

Por último éste método consigue reducir la tasa máxima de exposición, pero el número de items infrautilizados permanece invariable.

En cuanto a las restricciones en la selección de items, la investigación sobre TAIS se ha centrado en el estudio de sus propiedades psicométricas. Sin embargo, determinadas exigencias de índole aplicada obliga a plantear las restricciones en el algoritmo de selección de items para, por ejemplo asegurarse de que se aplican a todos los sujetos items con formato y contenido parecido.

Algunos TAIs establecen determinadas condiciones en el algoritmo de presentación, de tal forma que, por ejemplo, se seleccione el más informativo de los que comparten un formato determinado.

El problema es importante cuando las restricciones son muchas y variadas. Stocking y Swanson (1993) han propuesto el método de las desviaciones ponderadas, ideado para maximizar las propiedades deseable en una ubicación concreta. La idea fundamental de este modelo es que los especialistas establezcan las restricciones, y entonces formularlas matemáticamente de manera lineal, tratando de seleccionar como siguiente items aquel que maximiza la información o que al mismo tiempo satisface las restricciones, no estrictamente psicométrica, opuestas.

Las especificaciones del test o restricciones en la selección de items, puede ser de varios tipos:

1 Restricciones en las propiedades intrínsecas de items. Por ejemplo de

contenido, si es necesario cubrir varias áreas durante el tets. En el tipo o apariencia de los items, en la posición de la alternativa correcta.

2 De solapamiento, si se desea evitar que algunos grupos de items aparezcan juntos. Por ejemplo en el caso de que la presentación de un items proporcione la clave para responder a otro, items redundantes, etc.

3 Grupos de items que deben aparecer juntos, por ejemplo si todos ellos se refieren al mismo estímulo, como un párrafo de texto o una gráfica.

4 Restricciones de estadísticas relacionadas con la información que aportan los items para el nivel estimado de habilidad.

El método empleado para seleccionar los items de acuerdo con todas estas restricciones consiste en determinar a priori cual es el objetivo que se desea conseguir por ejemplo el numero de items máximo y mínimo que se desea aplicar de cada área y seleccionar en cada momento aquel items que en mayor medida concluya a alcanzar estos objetivos.

Para mejorar las condiciones de aplicación, una importante línea de trabajo sobre la investigación sobre TAI se ha orientado a establecer las conducciones de aplicación más confortables, intentando al mismo tiempo minimizar la repercusiones en la calidad de la medida. Se han ensayado condiciones que incrementan la sensación sujetiva de éxito en la prueba y otras en las que se permite la revisión de respuestas.

Sobre todo en contexto educativo se valoran muy positivamente quizás incluso más que la bondad de las medidas que un test aporte información para plantear objetivos instruccionales encaminados a superar las dificultades evaluadas. Algunas experiencias interesantes tienen que ver con el diseño de TAI estrechamente ligado a lo que podría considerarse un “entrenamiento adaptativo”.

Un problema que suele asociarse a los TAI es que, independientemente del nivel de habilidad de los sujetos, todos suelen tener una tasa de aciertos en torno al 50% además, la secuencia de presentación de los items se alejan bastante de los test de rendimiento, en los cuales la dificultad es progresiva. La cuestión es si estos aspectos singulares, asociados al procedimiento adaptativo de selección de items, tienen repercusiones en el estado motivacional con que algunos evaluados afrontan la prueba y por tanto también en su rendimiento. No todo es conseguir estimaciones precisas.

Mucho tiempo la investigación de TAI se hace entrada en el estudio de los algoritmos más eficaces y en las propiedades psicométricas de las estimaciones, las necesidades de índole aplicado, sobre todo en contextos educativos, impulsando una línea de trabajo sobre el tipo de errores que comete un alumno, para así proponer instrucciones adaptadas a los problemas.

Un buen ejemplo de este tipo de investigación es la que desarrollan Tatsuoka y sus colaboradores sobre lo que denominan método *regla-espacio* ,una estrategia orientada a evaluar y entrenar errores académicos relacionados con el procesamiento cognitivo y el conocimiento que intervienen en la resolución de fracciones.

4.-TAIs ON-LINE

1 Introducción

El modelado del alumno es un problema central en el diseño y desarrollo de un sistema tutor inteligente (STI). En efecto, si la característica que distingue a los STIs de los sistemas de Enseñanza Asistida por ordenador tradicionales es su capacidad de adaptación al alumno (Shute, 1995), el sistema debe ser capaz de determinar con la mayor precisión y rapidez posible cuál es su estado cognitivo, es decir, qué partes del dominio que pretendemos enseñarle son las que ya domina y cuáles son las que aún desconoce. Sólo de esta forma será posible adaptar el proceso instructor: saber qué estrategia instructora es más conveniente, qué acción recomendarle (estudio, resolución de ejercicio, juego), qué ejercicio se adecua más a su nivel de conocimiento, etc.

El problema del modelado del alumno puede dividirse en dos componentes:

(a) seleccionar la estructura de datos (*modelo del alumno*) que será usada para representar toda la información relativa al alumno: estado cognitivo, estrategias instructoras preferidas, pantallas visitadas, ejercicios resueltos, etc.; y

(b) elegir el procedimiento que utilizaremospara realizar el *diagnóstico*, es decir, para inferir dada la información generada en la interacción del alumno con el sistema (problemas resueltos, pantallas visitadas, etc.) el estado cognitivo del alumno. Evidentemente ambas componentes están íntimamente relacionadas, y por tanto lo ideal es diseñarlas y desarrollarlas simultáneamente. Una descripción más completa del problema se puede encontrar en (VanLehn, 1988) El diagnóstico es sin duda uno de los procesos más importantes dentro de cualquier STI, puesto que como ya hemos mencionado, de la calidad del modelo del alumno dependerá la capacidad de adaptación del sistema.

Desgraciadamente, no siempre se le presta la atención que merece, dado que el gran esfuerzo que supone desarrollar un sistema tutor inteligente hace que a menudo el problema del modelo del alumno se resuelva mediante la aplicación de heurísticos diseñados a tal fin. Pero la falta de consistencia de dichos heurísticos hace que el comportamiento del sistema sea impredecible, sobre todo en situaciones diferentes a las inicialmente previstas por sus diseñadores. Es por ello por lo que pensamos que, pese al esfuerzo adicional que supone, merece la pena utilizar teorías bien fundamentadas y ampliamente comprobadas que garanticen el funcionamiento óptimo del sistema en todas las situaciones posibles, y, en concreto, proponemos el uso de la *teoría de la probabilidad* como marco teórico. Además, queremos conectar el problema del modelado del alumno con la teoría de los *test adaptativos informatizados* (TAI) (Wainer, 1990), que se ha desarrollado dentro del campo de la psicometría y que pese a su capacidad demostrada para mejorar el proceso de diagnóstico tanto en precisión como en tiempo (Huang, 1996) no ha sido aún utilizada dentro del campo de los STIs. Este artículo se estructura de la siguiente forma: en la siguiente sección describimos los conceptos básicos de los test adaptativos informatizados, así como el sistema SIETTE, que es una herramienta web basada en la teoría de la

respuesta al ítem unidimensional que cumple con dos objetivos distintos: (a) permite que los profesores definan de una forma muy sencilla un test adaptativo informatizado; y (b) permite que los alumnos realicen los test definidos y sean evaluados por el sistema, todo ello a través de la web (Ríos, Millán et al., 1999). Posteriormente, describimos un enfoque basado en redes bayesianas (Pearl, 1988) que permite realizar test adaptativos en los que se mide más de una habilidad, más adecuados si el objetivo del test no es meramente evaluar al alumno sino llevar a cabo el proceso de diagnóstico en un STI. Finalmente, presentamos las conclusiones obtenidas en la realización de ambos trabajos y las líneas futuras de investigación.

2 Diagnóstico basado en el modelo TRI

2.1 Tests adaptativos informatizados

El uso de los tests para la evaluación es una técnica ampliamente usada en el campo de la educación. Los métodos tradicionales de diseño y administración de tests dependían en gran medida de que éstos fuesen orientados a un individuo o a un grupo. Los tests administrados a grupos son menos costosos en tiempo y recursos que los individuales y además tienen la ventaja de que todos los examinandos están en igualdad de condiciones.

Como contrapartida, los tests de este tipo deben contener ítems con tantos niveles de dificultad como posibles niveles de conocimientos puedan existir en el grupo de alumnos que va a realizarlos, mientras que los tests administrados individualmente contienen ítems elegidos de forma más apropiada. Este hecho puede acarrear consecuencias no deseables como el aburrimiento de alumnos con niveles altos de conocimiento o el desconcierto y la frustración en los alumnos menos aventajados. A principios de los 70 surgieron trabajos que apuntaban que el uso de tests más flexibles aliviaría en parte estos problemas. En (Lord, 1970) se establece la estructura teórica de un test de administración masiva pero adaptado individualmente: "*la idea básica de un test adaptativo es imitar lo que un examinador sensato haría"* (Wainer & Messick, 1983), es decir, si un examinador hace una pregunta que resulta ser demasiado difícil, la siguiente debería ser más fácil. Sin embargo, probar los tests adaptativos de una forma seria no fue posible hasta principios de los 80, con la aparición de ordenadores potentes y menos costosos. Surgen entonces los llamados test adaptativos informatizados (TAI). Un Test adaptativo informatizado es básicamente un test administrado por ordenador donde la presentación de cada ítem y la decisión de finalizar el test se toman de forma dinámica basándose en la respuesta del alumno y en la estimación de su nivel de conocimiento. En términos más precisos, un TAI es un algoritmo iterativo que que comienza con una estimación inicial del nivel de conocimiento del alumno y que tiene los siguientes pasos: (1) Todas las preguntas que no se han administrado todavía son examinadas para determinar cuál será la mejor para ser propuesta a continuación, según el nivel de conocimiento estimado del alumno; (2) la pregunta es planteada y el alumno responde; (3) de acuerdo con la respuesta del alumno, se realiza una nueva estimación de su nivel de conocimiento. Los pasos del 1 al 3 se repiten hasta que se cumpla alguno de los criterios de terminación definidos. Los TAIs tienen importantes ventajas frente a los tests tradicionales a lápiz y papel entre las que destacan: (a) decremento significativo en la longitud de los tests; (b) estimaciones más precisas del nivel de conocimiento del alumno; (c) mejora en la motivación de los alumnos; (d) se puede almacenar un gran banco de

preguntas, incluyendo enunciados y posibles respuestas con contenido multimedia. Los elementos básicos de un TAI son:

- Modelo de respuesta del ítem. Este modelo describe como el sujeto responde al ítem según su nivel de conocimiento. Cuando se llevan a cabo mediciones del nivel de conocimiento, cabe esperar que el resultado obtenido no dependa del instrumento utilizado, es decir, la medida ha de ser invariante con respecto al tipo de test y al sujeto al que se le aplica el test.

- Banco de preguntas. Constituye uno de los elementos fundamentales para la creación de un TAI. Para definir un banco de preguntas eficiente se deben especificar las distintas áreas de conocimiento del dominio. Una vez hechas las especificaciones del contenido del test, el banco de preguntas debe contener ítems en suficiente número, variedad y niveles de dificultad (Flaugher,1990).

- Nivel de conocimiento de entrada. Elegir de forma adecuada el nivel de dificultad de la primera pregunta que se realice en el test puede reducir sensiblemente la longitud del mismo. Para ello se pueden usar diferentes criterios como tomar el nivel medio de los sujetos que han realizado el test previamente, o crear un perfil de sujeto y usar el nivel medio de los alumnos con un perfil similar (Thissen & Mislevy, 1990).

- Método de selección de preguntas. Un test adaptativo selecciona el siguiente ítem que va a ser presentado en cada momento en función del nivel estimado del conocimiento del alumno y de las respuestas a los ítems previamente administrados. Seleccionar el mejor ítem puede mejorar la precisión en la estimación del nivel de conocimiento y reducir la longitud del test.

- Criterio de terminación. Para decidir cuándo debe finalizar un test se pueden usar diferentes criterios tales como parar cuando se haya alcanzado una precisión determinada en la medida del nivel de conocimiento, cuando se hayan planteado un número determinado de ítems, etc.

2.2 Teoría de respuesta al ítem

La mayor parte de las aplicaciones prácticas de la teoría de la medida en Psicología y Educación están basadas en la Teoría Clásica de Tests (TCT), cuyas deficiencias alentaron la búsqueda de modelos alternativos. Entre los que mayor difusión han tenido destacan los basados en *la Teoría de la respuesta al ítem* (TRI) (Lord, 1968), (Hambleton, 1989), inicialmente conocida como *teoría del rasgo latente*. La TRI, partiendo de hipótesis restrictivas, intenta dar fundamentos probabilísticos al problema de la medición de rasgos no observables. Su nombre es debido a que se consideran los ítems como las unidades básicas de los tests. Todos los modelos TRI tienen unas características comunes: (a) suponen la existencia de rasgos o aptitudes latentes que permiten predecir o explicar la conducta de un examinando ante un ítem de un test; (b) la relación entre el rasgo y la respuesta del sujeto al ítem puede describirse por medio de una función monótona creciente, denominada Curva característica del item (CCI). Los primeros modelos

aparecidos son conocidos con el nombre de ``modelos normales" ya que la forma de la ICC era la de una distribución normal (Lord, 1968). Las dificultades para el manejo analítico de esta función llevaron a los *modelos logísticos*, basados en la función de distribución logística, entre los que destacan los de un parámetro (Rasch, 1960), y los de dos y tres parámetros (Birnbaum, 1968). Todos estos modelos están basados en la suposición de independencia local que afirma que si la aptitud θ que explica el rendimiento en el test permanece constante, las respuestas de los examinandos a un par de ítems cualquiera, son estadísticamente independientes.

Por otra parte, el *método bayesiano* calcula el nivel de conocimiento para el que la distribución aposteriori es máxima. Esta distribución es proporcional al producto de la función de probabilidad y la función de densidad a priori, es decir, $P(\theta/u) = L(\theta/u)\, f(\theta)$.
En cuanto a los métodos de selección los más comunes son el de la *máxima información* (Weiss & Kingsbury, 1984), que consiste en seleccionar el ítem que haga máxima la información del ítem para el nivel de conocimiento estimado hasta el momento, y los *bayesianos*, como el de Owen (Owen, 75), que selecciona la pregunta que hace mínima la varianza a posteriori de la distribución del conocimiento.

2.3 El sistema SIETTE

El sistema SIETTE (Ríos, Millán et al., 1999) se basa en una implementación discreta de la teoría de test adaptativos, a la que se le han añadido algunos nuevos elementos para mejorar su funcionalidad y para la que se ha utilizando la WWW como interfaz de desarrollo y realización de test. La arquitectura general del sistema SIETTE se muestra consta de dos módulos claramente diferenciados. El primer módulo ha sido diseñado para que un conjunto de profesores pueda insertar preguntas y definir los tests a realizar. Además de las preguntas, el profesor puede definir un conjunto de temas en los que se divide la matera, organizar las preguntas de acuerdo a estos temas, definir el numero de respuestas posibles a mostrar al alumno para cada pregunta, así como los parámetros propios de la composición de cada tests, como el porcentaje de preguntas de cada tema que en el intervienen, el modo de selección de preguntas y el criterio de finalización, el número mínimo y máximo de preguntas a realizar, etc. El módulo de generación de tests es al que accede el alumno para realizar las pruebas, las preguntas se generan de forma individualizada para cada alumno según la materia y el test que haya elegido. Se mantiene un registro temporal de la evolución del alumno durante la sesión que se tiene en cuenta en el proceso de selección de preguntas y un registro histórico de las respuestas a cada pregunta que servirá como fuente de información para el aprendizaje automático de los parámetros de las preguntas. Se ha incluido en el sistema un proceso automático de validación y activación de las preguntas y de los tests diseñados por los profesores que permite, por un lado, separar las tareas de edición y realización de tests, eliminando los posibles problemas de inconsistencia en el caso de edición y realización simultánea; y, por otro lado, se encarga de comprobar que la definición de las preguntas y de los tests es coherente (por ejemplo, si existe un número suficiente de preguntas de cada tema, si se han incluido suficientes respuestas alternativas para cada pregunta, etc.) La estructura del sistema SIETTE se basa en los conceptos tradicionales de asignatura y tema. SIETTE puede trabajar de forma simultánea con varias materias independientes. El acceso al sistema se efectúa mediante una clave asociada a cada asignatura. Esta clave puede ser

personal o compartida por un conjunto de profesores de dicha materia. Cada asignatura o materia se subdivide en temas en la forma que el profesor crea más adecuada, usando para ello el editor para añadir, modificar o eliminar temas. Los temas representan grandes bloques de la asignatura y no necesariamente conceptos concretos. En su versión actual SIETTE no maneja ninguna información sobre la interdependencia de los temas. Una vez definidos los temas, el profesor debe definir las preguntas o cuestiones que compondrán los tests. El sistema hasido diseñado de forma que también almacene una posible ayuda, asociada a cada posible respuesta del alumno, o en caso de que éste requiera mayor información para resolver la pregunta. Las preguntas y las respuestas se pueden introducir como texto simple al que, si así se desea, se puede añadir de código HTML, JavaScript, *applets* y metacódigo PHP. Por consiguiente el formato de las cuestiones admite cualquier objeto multimedia. nAdemás del enunciado, las respuestas y la ayuda, el profesor debe proporcionar algunos datos sobre la cuestión como, por ejemplo, el número de alternativas incorrectas a mostrar, el grado de dificultad y el factor de discriminación de la pregunta, la distribución en pantalla, etc. y debe asociar cada pregunta a uno o varios temas de entre los definidos anteriormente. Dado que el enunciado y las respuestas permiten el uso de metacódigo y applets, cada cuestión introducida puede ser en realidad un esquema generador de cuestiones, lo que permite una gran variedad de preguntas. También es posible simular como quedará la pregunta al presentarla. Esto es especialmente útil en el caso de preguntas generativas.

2.4 Edición de tests

Una vez definidas las preguntas de una materia y asociadas cada una de ellas a uno o varios temas, el profesor puede definir los tests que se van a realizar. Un test se compone de un conjunto de preguntas seleccionadas según diversos criterios basados en la teoría de tests adaptativos. Al definir el test deben seleccionarse los criterios que se usarán tanto para la selección de preguntas como para la finalización del test.

A diferencia de los sistemas clásicos en TRI/TAI, y a fin de ajustar la composición de los tests, el profesor puede especificar los porcentajes de preguntas de cada tema que compondrán cada test, y el número mínimo de preguntas de cada uno de ellos. Esto debe garantizar que un alumno que supere el test tiene un conocimiento suficientemente homogéneo de la materia. También por cuestiones prácticas se fija un número mínimo y máximo de preguntas para garantizar que en cualquier caso el test tiene un final, se alcance o no el criterio de finalización estadístico. el algoritmo para la generación de tests es el expuesto más arriba: (a) selección de la pregunta: que puede ser aleatoria, adaptativa o por máxima probabilidad, y bayesiana o por mínima desviación tipica a posteriori. También se tienen en cuenta la distribución de temas en la composición de los tests que ha indicado el profesor y aspectos funcionales tales como la no repetición de preguntas en un mismo tests o en sucesivos tests realizados por el mismo alumno; (b) estimación del nivel de conocimiento del alumno; y (c) criterio de finalización: según las especificaciones del profesor y atendiendo a la desviación típica de la variable estimada, la cota del error sobre un determinado nivel y los mínimos y máximos de preguntas previamente configurados. A diferencia de la teoría clásica de la TRI/TAI, que trabaja con funciones definidas sobre el conjunto de los números reales, SIETTE emplea una aproximación numérica a estas

funciones. Esto se traduce en una mayor facilidad para la aplicación de los métodos bayesianos ya que no es necesaria la resolución exacta de las ecuaciones. Igualmente desde el punto de vista computacional el proceso es más eficiente si el número de intervalos considerados es pequeño. Por otra parte, el uso de aproximaciones numéricas tiene la ventaja de que no está restringido a ninguna familia de funciones concreta para definir las ICC de las cuestiones. Si bien se ha seguido utilizando la función logística como base para la definición de estas curvas de forma paramétrica por parte del profesor, nada impide utilizar otras distribuciones que no se ajusten a esta familia de curvas, con lo que el ajuste a la distribución real de dificultad de las preguntas puede ser mejor. Este ajuste puede llevarse a cabo mediante técnicas de aprendizaje estadístico de forma directa (Conejo, Millán et al., 2000), sin necesidad de estimaciones restringidas a una familia concreta de curvas. SIETTE permite el acceso tanto de forma identificada como anónima. En el caso de usuarios registrados, el sistema tiene en cuenta los tests realizados anteriormente por el alumno, y es capaz de mantener el modelo temporal del alumno como punto de partida para una nueva evaluación. El alumno puede seleccionar el test a realizar de entre todos los test disponibles y puede configurar ciertos parámetros, como la presentación de las respuestas correctas inmediatamente después de la resolución de cada pregunta o sólo al final del test

3 Diagnóstico mediante redes bayesianas

En esta sección vamos a describir cómo las redes bayesianas pueden ser utilizadas en el problema de diagnóstico del alumno. Para ello definimos en primer lugar el modelo estructural que servirá como soporte del proceso evaluador (nodos, enlaces y parámetros), y después presentamos los resultados obtenidos en la evaluación del modelo propuesto. Dicha evaluación ha sido realizada utilizando alumnos simulados.

3.1 Modelo estructural

Para utilizar redes bayesianas en el problema de diagnóstico, lo primero que tenemos que hacer es definir los elementos básicos: variables, enlaces entre ellas y parámetros. A continuación presentamos el modelo estructural integrado que hemos desarrollado, que no sólo permite realizar el diagnóstico a diferentes niveles de granularidad, sino que propone simplificaciones notables para la especificación de los parámetros. Nos centraremos en el diagnóstico basado en preguntas tipo test, aunque en principio sería posible considerar cualquier tipo de preguntas siempre que el sistema tuviera la capacidad de comprobar si la solución propuesta por el alumno es o no correcta.

3.1.1 Variables

Consideraremos dos tipos básicos de variables: variables para medir el grado de conocimiento alcanzado por el alumno, y variables para recolectar evidencia. A su vez, y para una evaluación más detallada, las variables de conocimiento se definen a diferentes niveles de granularidad. Describimos a continuación cada uno de estos tipos. Variables para medir el conocimiento del alumno. Vamos a utilizar tres niveles de granularidad, que creemos que serán suficientes en la mayoría de las aplicaciones, pero como veremos no hay problema alguno en modelar más niveles utilizando el mismo enfoque. En el nivel inferior aparecen los *conceptos*, que representan las unidades mínimas en las que se puede

descomponer el conocimiento. El nivel inmediatamente anterior contiene los *temas*, que son agrupaciones de conceptos. Por último, aparecen las *asignaturas*, que son agrupaciones de temas. Consideraremos que todos los nodos son binarios, pero la interpretación que se da a la probabilidad de los distintos tipos de nodo es diferente: en los nodos concepto, representan la probabilidad de que el concepto se conozca o no se conozca, mientras que en los nodos asignatura y tema dicha probabilidad se interpreta como una medida del grado de conocimiento alcanzado en el tema y la asignatura. La justificación teórica que permite considerar las probabilidades de la forma descrita aparece en (Millán, Pérez-de-la-Cruz et al., 2000). Variables para recolectar evidencia. En nuestro caso serán preguntas tipo test multirespuesta. Las respuestas a dichas preguntas pueden ser correctas o incorrectas.

3.1.2 Enlaces

En esta sección vamos a definir las relaciones que se establecen entre las variables definidas. Respecto a las relaciones entre variables para medir el conocimiento, consideraremos que dominar un nodo de conocimiento tiene influencia causal en dominar aquellos nodos de conocimiento del nivel inmediatamente anterior en la jerarquía de granularidad que estén con él relacionados. En cuanto a la relación entre los nodos de conocimiento y las preguntas, consideraremos que poseer el conocimiento tiene influencia causal en responder adecuadamente a las preguntas.

La red bayesiana se divide en dos partes que se solapan en los conceptos: la parte que soporta el proceso de diagnóstico, en el que se determina a partir de las respuestas del alumno el conjunto de conceptos que conoce/no conoce, y la parte que soporta el proceso de evaluación, en el que a partir de los resultados obtenidos en el proceso anterior se determina una medida del grado de conocimiento alcanzado por el alumno, tanto en la asignatura como en cada uno de los temas de los que consta. Cada una de estas partes se modela con un tipo de red bayesiana diferente: la parte de evaluación se modela con una red bayesiana clásica, mientras que para la parte de diagnóstico se utiliza una red bayesiana dinámica, puesto que en este caso es claro que los nodos tipo evidencia cambian con el tiempo, es decir, el hecho de que un alumno conteste correctamente a una pregunta relacionada con ciertos conceptos no quiere decir que siempre que le planteásemos una pregunta relacionada con tales conceptos la contestase también correctamente.

3.2 Parámetros

Definidas las relaciones, los parámetros que necesitamos especificar son:
Probabilidades a priori de los nodos concepto. Para ello podemos utilizar la información que haya disponible sobre el alumno en cuestión. En ausencia de información utilizaremos la distribución uniforme, es decir, consideraremos igualmente probable que domine el concepto o que no lo domine. Probabilidades condicionadas de los temas dados los conceptos, y de la asignatura dados los conceptos.

4 Conclusiones y trabajo futuro

En las páginas precedentes hemos expuesto una visión panorámica de los trabajos teóricos y prácticos realizados por nuestro grupo con la finalidad de proporcionar herramientas de modelado robustas y bien fundamentadas en el campo de de los STI. Creemos haber mostrado que tanto la teoría clásica de la TRI como las más recientes redes bayesianas son un punto de partida adecuado para ello. Sin embargo, aún queda un largo camino por recorrer. Por ejemplo, será necesario integrar más estrechamente ambos enfoques para obtener una herramienta más versátil. Por otra parte, la evidencia que se tiene en cuenta en las versiones actuales de nuestros sistemas es muy reducida: únicamente las respuestas del alumno a las preguntas multirrespuesta. Para integrar realmente estas herramientas en un STI, será necesario tomar además en consideración otro tipo de datos, como pueden ser peticiones de ayuda del alumno y, en general, cualquier tipo de *episodios instructivos*. En esta línea se desarrollan nuestras actuales investigaciones.

5.- BIBLIOGRAFIA

Olea, J. y Ponsoda, V. (1996). Tests adaptativos informatizados. En J. Muñiz (Coor.): Psicometría.Cap. 15, p.p. 729-783 Madrid: Ed. Universitas S.A.

Renom, J. y Doval, E. (1999), Test Adaptativos Informatizados. En J. Olea, V. Ponsoda y G. Prieto (Eds), Tests informatizados. Fundamentos y Aplicaciones. Cap 6, p.p. 127-162. Madrid: Pirámide.