SIETTE: Sistema Inteligente de Evaluación mediante Test para TeleEducación

Ricardo Conejo; Eduardo Guzmán
E.T.S. Ing. Informática
Universidad de Málaga

1. INTRODUCCIÓN

La evaluación ha sido siempre una parte importante del proceso de enseñanza y aprendizaje. Por una parte, los profesores necesitan conocer cual es el conocimiento que han adquirido los alumnos tras el proceso de enseñanza para actuar en consecuencia, y por otra, los propios alumnos necesitan contrastar de manera no subjetiva el conocimiento adquirido. De igual manera en el campo de la Enseñanza Asistida por Ordenador, EAO y de los Sistemas Instructores Inteligentes, STI, especialmente en estos últimos, la evaluación ha sido siempre una parte importante del sistema.

Uno de los mecanismos de evaluación más extendidos, por su facilidad de corrección, es la realización de tests. La realización de tests tiene la ventaja de sistematizar la evaluación, por lo que ha sido ampliamente usada en aplicaciones de enseñanza asistida por ordenador y en sistemas tutores inteligentes (Brusilovsky 1999). En los sistemas de enseñanza tradicional, el uso de tests de papel y lápiz también tiene sus inconvenientes: las preguntas del test son en general las mismas para todos los alumnos, el número de preguntas tampoco varia, el tipo de preguntas se limita a una o varias opciones, o a una respuesta corta, etc. Por otra parte la evaluación que se obtiene tras la realización de un test mediante papel y lápiz raramente tiene en cuenta mas que el número de respuestas acertadas, y no suelen tener en cuenta la variabilidad en la dificultad de las preguntas, ni otros factores como la probabilidad de acertar una pregunta al azar, sin realmente saber la respuesta.

En el campo de la psicometría, en el que la realización de tests ha sido siempre tradicional, surge en los años 1960-1970 la teoría de la respuesta al ítem con el objeto de mejorar la evaluación y cuantificar de manera probabilista los resultados del test y los posibles errores de evaluación debidos al azar. En la teoría de respuesta al ítem se asume que el conocimiento del alumno puede medirse mediante una única variable denominada *rasgo*. A partir de las respuestas del alumno a la secuencia de preguntas del test y mediante un procedimiento estadístico que tiene en cuenta la dificultad de las preguntas se obtiene un estimador de este rasgo, a partir de la función de distribución de las probabilidades de que el conocimiento real del alumno tome determinado valor.

La teoría de tests adaptativos informatizados, TAI, evoluciona a principios de los años 1980s, en general basándose en la aplicación mediante ordenador de teoría de respuesta al ítem. Básicamente el objeto de esta nueva teoría es mejorar la evaluación mediante dos mecanismos: la selección de las preguntas mas adecuadas a cada alumno según su nivel de conocimiento estimado en cada momento y la realización del mínimo número de preguntas necesarias para garantizar una cota superior del error de estimación.

Por otra parte, en la última década, los sistemas de enseñanza asistida por ordenador y los sistemas tutores inteligentes han hallado en Internet, y más concretamente en la World Wide Web, WWW un medio ideal de difusión. De hecho, ya son muchos los centros de enseñanza y universidades que incluyen en sus portales al menos el soporte para la difusión de contenidos y material docente. Desde el punto de vista de los sistemas tutores inteligentes la WWW ofrece muchas ventajas en comparación con otros medios más tradicionales de difusión de programas: no requiere de instalaciones adicionales al usuario; se puede acceder al sistema desde cualquier lugar sin necesidad de transportar el software; los contenidos se distribuyen y mantienen actualizados de forma inmediata; facilita la construcción de sistemas distribuidos; existe la posibilidad de obtener datos sobre la efectividad de la instrucción y consiguientemente mejorarla, etc. Sin embargo, uno de los inconvenientes del uso de la WWW es la complejidad técnica que requiere tanto el desarrollo de sistemas personalizados como la creación de contenidos.

SIETTE es una aplicación eficiente de la teoría de test adaptativos informatizados y de la teoría de respuesta al ítem para la realización de tests a través de WWW. SIETTE aporta técnicas propias para sacar el máximo provecho de las características de este nuevo medio, ofreciendo un valor añadido a la evaluación mediante test. SIETTE incluye además una herramienta de autor para que los profesores puedan crear los tests también a través de Internet, y de un simulador para estudiar el funcionamiento de los tests en condiciones controladas, estudiar su efectividad, y calibrar los parámetros de un conjunto de preguntas. SIETTE ha sido diseñado como un componente autónomo, aunque puede también integrarse como parte de un sistema tutor inteligente. En concreto se prevé su integración en la arquitectura MEDEA

La primera versión de SIETTE se desarrolló en 1998 (Ríos, 1998). En la actualidad se está desarrollando una segunda versión que incluye tanto mejoras técnicas como nuevas funcionalidades. En este artículo se exponen tanto el sistema actualmente implementado y en uso, como las nuevas funciones que integrarán esta segunda versión.

En el siguiente epígrafe se presenta la arquitectura general del sistema SIETTE, sus principales componentes y funciones, incluyendo algunas pantallas de ejemplo. Seguidamente en los apartados 3, y 4 se introducen los fundamentos teóricos de SIETTE: la teoría de respuesta al ítem y la teoría de test adaptativos describiendo las características propias de la implementación en SIETTE respecto a la teoría clásica. El apartado 5 describe los estudios de simulación sobre los algoritmos y las características propias implementadas en SIETTE. El apartado 6 está dedicado al problema de la estimación de los parámetros de los ítems, y en él se expone brevemente la solución adoptada en SIETTE, así como las estrategias de aprendizaje automático integradas en el sistema. El apartado 7 analiza en detalle los distintos tipos de ítems que pueden emplearse para formular cuestiones, y cómo se integran en el proceso de inferencia de SIETTE. El apartado 8 trata sobre la integración de SIETTE como módulo de evaluación de un sistema tutor inteligente, haciendo especial énfasis en los mecanismos de evaluación simultanea de múltiples conceptos. Finalmente el apartado 9 presenta las conclusiones, describiendo las características de la implementación actual y las líneas de investigación en marcha.

2. ARQUITECTURA DEL SISTEMA SIETTE

SIETTE es un sistema de evaluación mediante tests adaptativos basado en la teoría de respuesta al ítem, que ha sido diseñado para ser usado sobre la World Wide Web. Utilizando un navegador estándar como interfaz gráfica de usuario, ofrece a los alumnos un aula virtual en el que realizar tests y permite a los profesores crear nuevos tests o añadir nuevas cuestiones a los tests ya creados.

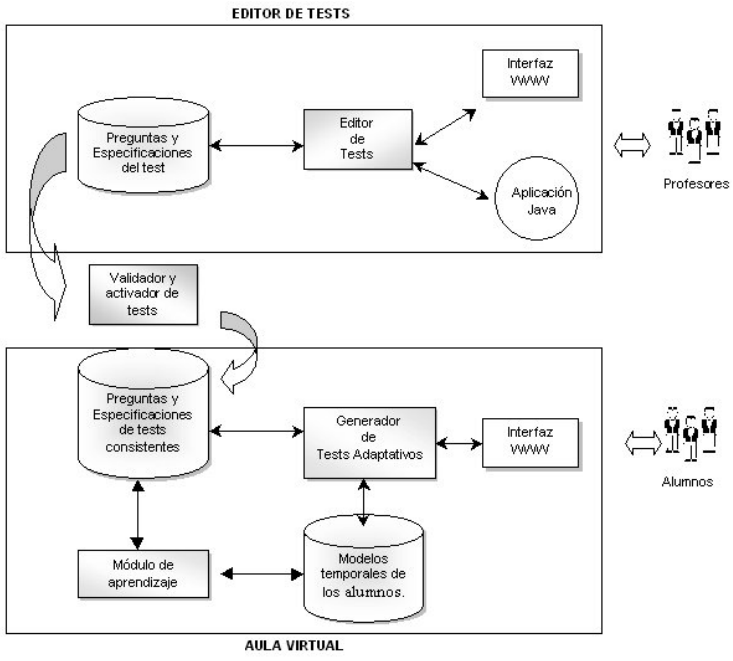

Figura 1: Arquitectura del sistema SIETTE.

La arquitectura del sistema SIETTE, contiene los principales componentes de un test adaptativo agrupados en seis módulos principales, su representación gráfica se muestra en la figura 1:

- La *base de conocimientos*, Está compuesta por *el curriculum o estructura del temario, las especificaciones de tests* y el *banco de ítems,* que constituye la colección de posibles cuestiones a presentar en un test, todas ellas calibradas con una serie de parámetros.
- El *generador de tests*. Es el módulo principal del sistema SIETTE. Es el encargado de seleccionar las preguntas a plantear al alumno, según las especificaciones del test y del *modelo temporal del alumno*. En los apartados 3 y 4 se exponen los fundamentos teóricos del algoritmo de evaluación utilizado.
- El *módulo de edición* permite a los profesores acceder a la base de conocimientos para almacenar las preguntas y respuestas y especificar el *currículum,* y *los tests* sobre los temas que se desea evaluar.
- *El módulo de comprobación y activación.* Para que los tests diseñados por el profesor pasen a disposición de los alumnos, sus datos deben ser validados. Sólo aquellas especificaciones de tests que cumplen los criterios mínimos de consistencia requeridos serán activadas por este *módulo*.
- El *módulo de aprendizaje* que se encarga de realizar una calibración de los parámetros de los ítems a partir de la información obtenida de forma empírica tras las sucesivas ejecuciones de los tests por parte de los estudiantes. Esta calibración es necesaria ya que los valores asociados a los ítems que introducen por los profesores, sólo se consideran aproximaciones iniciales. (véase el apartado 6)

Además de los componentes que han sido descritos anteriormente, SIETTE dispone de un simulador que permite el estudio empírico de las variables estadísticas que se manejan, así como de la influencia de los parámetros de los ítems, e incluye mecanismos para el aprendizaje automático de estos parámetros. (véase epígrafe 5)

La base de conocimientos

SIETTE es un sistema genérico para evaluación. Está estructurado por materias o asignaturas independientes. Cada una de estas asignaturas tiene una base de conocimientos distinta. La base de conocimientos de SIETTE está formada por tres tipos de objetos: Los *conceptos*, que son los temas o elementos en los que se descompone la asignatura. Están estructurados jerárquicamente, formando un *curriculum*. Los *tests*, que representan las sesiones de evaluación. Cada *test* está formado por un conjunto de *ítems* o cuestiones. Los *ítems* están asociados a uno o varios conceptos o temas. La definición de los tests se hace en función de los temas sobre los que se desea evaluar.

En el apartado 8 se explica la relación existente entre *temas*, *ítems* y *tests*. Por el momento y a efectos de la exposición que se hace en los siguientes apartados puede suponerse simplemente que el curriculum contiene un sólo tema al que están asociadas todas las cuestiones y que existe un único test sobre este tema que utiliza estas cuestiones.

El generador de tests

Una vez que el profesor ha elaborado un test, los alumnos pueden acceder a él a través de una interfaz web denominada *aula de tests*. Las preguntas se generan dinámicamente y de forma individualizada para cada alumno según la materia y el test que haya elegido.

El primer paso que tiene que dar el alumno es identificarse dentro del sistema. SIETTE permite el acceso tanto de forma identificada como anónima. Para la primera opción, el usuario deberá introducir un identificador personal y una contraseña. En caso de que sea la primera vez que el usuario accede al sistema, deberá dar sus datos personales: nombre, apellidos, correo electrónico, así como seleccionar un identificador personal y una contraseña. Una vez terminado el proceso de identificación, se le muestran los tests disponibles para él. Cuando el usuario selecciona el test que quiere realizar, SIETTE le mostrará información sobre este test antes de empezar a formularle cuestiones. Si posteriormente quisiera volver a realizar otro o el mismo test que resolvió anteriormente, únicamente tendrá que introducir en el sistema su identificador y palabra de paso, puesto que ya consta como usuario registrado.

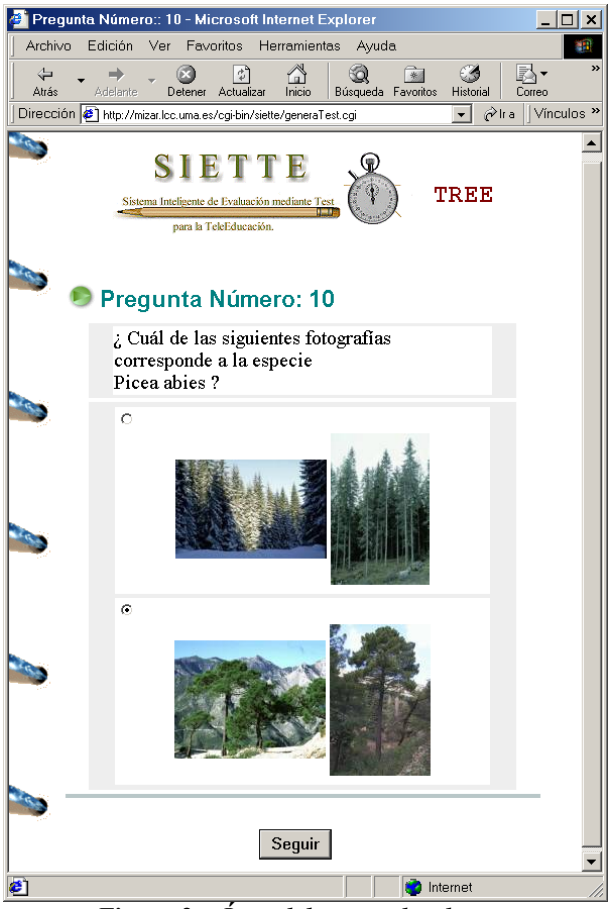

Figura 2a: Ítem del generador de tests

Figura 2b: Corrección de un ítem

Para los usuarios registrados el sistema tiene en cuenta los tests realizados anteriormente, y es capaz de mantener el modelo temporal como punto de partida para una nueva evaluación. Durante el proceso de realización del test, el sistema mantiene un registro temporal de la evolución del alumno, que se tiene en cuenta en el proceso de selección del siguiente ítem; y un registro histórico de las respuestas a cada cuestión, que servirá como fuente de información para el aprendizaje automático de los parámetros de las preguntas.

Las figuras 2a y 2b muestran un estado intermedio durante la realización de un test. Ambas corresponden a páginas generadas dinámicamente por el generador de tests de SIETTE para un test de botánica incluido en el sistema TREE (Rios, 99). Se muestran al alumno dos pares de fotografías, dos de ellas correspondientes a la especie *Picea abies* y otras dos de la especie *Pinus pinaster*, y se pide al alumno que seleccione cuales corresponden a la *Picea abies*. Una vez que el alumno responde a una pregunta, si se ha previamente seleccionado esta opción al comienzo del test, el generador indicará cuál es la solución correcta. En este caso, como se puede apreciar, el alumno no ha respondido correctamente. El generador muestra su respuesta con una marca en rojo y la respuesta correcta con otra marca en verde.

Una vez finalizado el test, el sistema devolverá (figura 3) el nivel estimado de conocimiento del alumno, el número de preguntas que ha realizado, así como el número de preguntas que han sido respondidas correctamente. Alternativamente, si SIETTE se usa integrado en un sistema instructor inteligente, al terminar la evaluación esta información pasa al sistema tutor mediante la llamada a una URL distinta para cada nivel de conocimientos, (Arroyo,2001), o mediante una misma URL a la que se pasan estos datos mediante parámetros.

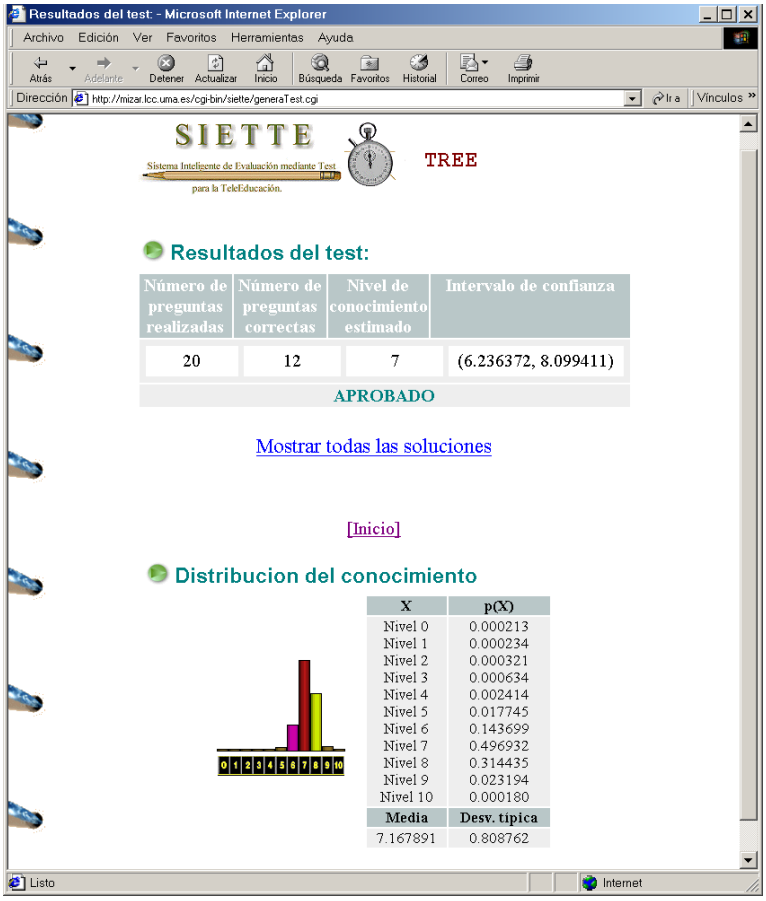

Figura 3: *Evaluación final de un test.*

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.

Los módulos de edición.

El editor de tests es una herramienta de autor que permite a los profesores insertar cuestiones, y definir los tests que pueden realizar los alumnos. La información generada por el editor se almacena en una base de conocimiento temporal hasta que el módulo de validación y activación las transfiera a la base de conocimiento del generador de tests, estando a partir de este momento a disposición de los estudiantes.

Actualmente coexisten dos editores. Un editor a través de la web, que fue desarrollado en la primera versión de SIETTE, y una nueva herramienta para edición y gestión de tests desarrollada en Java, que no requiere el uso del navegador.

La figura 6 muestra la interfaz del editor web. Para acceder a ella el profesor debe disponer de un identificador y una contraseña válida para la asignatura que desee modificar Cada asignatura o materia se subdivide en temas en la forma que el profesor crea mas adecuada, usando para ello las secciones para añadir, modificar o eliminar temas. Los temas representan grandes bloques de la asignatura sobre los que se desea evaluar y no necesariamente conceptos concretos. Una vez definidos los temas, el profesor definirá los ítems o cuestiones que compondrán los tests. Para ello rellena una ficha en la que se incluye el enunciado, la respuesta correcta y una o varias alternativas de respuestas incorrectas. El sistema ha sido diseñado de forma que también almacene una posible ayuda, asociada al enunciado del ítem, en caso de que el alumno solicite más información para resolver la pregunta. También es posible añadir un refuerzo por cada una de las respuestas, de forma que si el alumno no ha acertado al resolver la pregunta, el sistema puede indicarle por qué su respuesta no es válida. Esta característica convierte a SIETTE en un sistema no sólo evaluador sino también en cierto modo instructor.

Los textos de las preguntas y las respuestas son trozos de código HTML, a los que se puede añadir JavaScript, Applets de JAVA, e incluso metacódigo en PHP, PERL, JSP, etc.. El formato de las cuestiones admite cualquier objeto multimedia. Además del enunciado, las respuestas, la ayuda, y los refuerzos, el profesor debe proporcionar algunos datos sobre la cuestión, como por ejemplo, el número de alternativas incorrectas a mostrar, el factor de dificultad y debe asociar cada pregunta a uno o varios temas de entre los definidos anteriormente. Opcionalmente, puede proporcionar otros parámetros como el factor de adivinanza, el factor de discriminación, la distribución en pantalla, etc. 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. Es posible simular como quedará la pregunta al presentarla. Esto es especialmente útil en el caso de preguntas generativas.

Figura 4. *Interfaz web para edición de tests, temas e ítems.*

Una vez definidas las preguntas de una materia, 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 tanto para la selección de preguntas como para la finalización del test. 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. (véase apartado 4)

La nueva herramienta de edición mantiene las características ofrecidas por el anterior editor y añade otras nuevas. Permite dos modos de trabajo, según exista conexión permanente a Internet durante el proceso de modificación del test, o la conexión se realice únicamente al final para llevar a cabo la transferencia de las nuevas especificaciones de los tests al sistema SIETTE.

En la figura 5 se muestra el aspecto que ofrece la nueva herramienta de gestión de tests. Como se puede apreciar, esta herramienta muestra la estructura del currículum de la asignatura en forma de árbol. Para realizar una modificación o nueva inserción de un tema o de una cuestión, el profesor sólo tiene que seleccionar el elemento con el botón de la derecha del ratón y elegir la opción que desee.

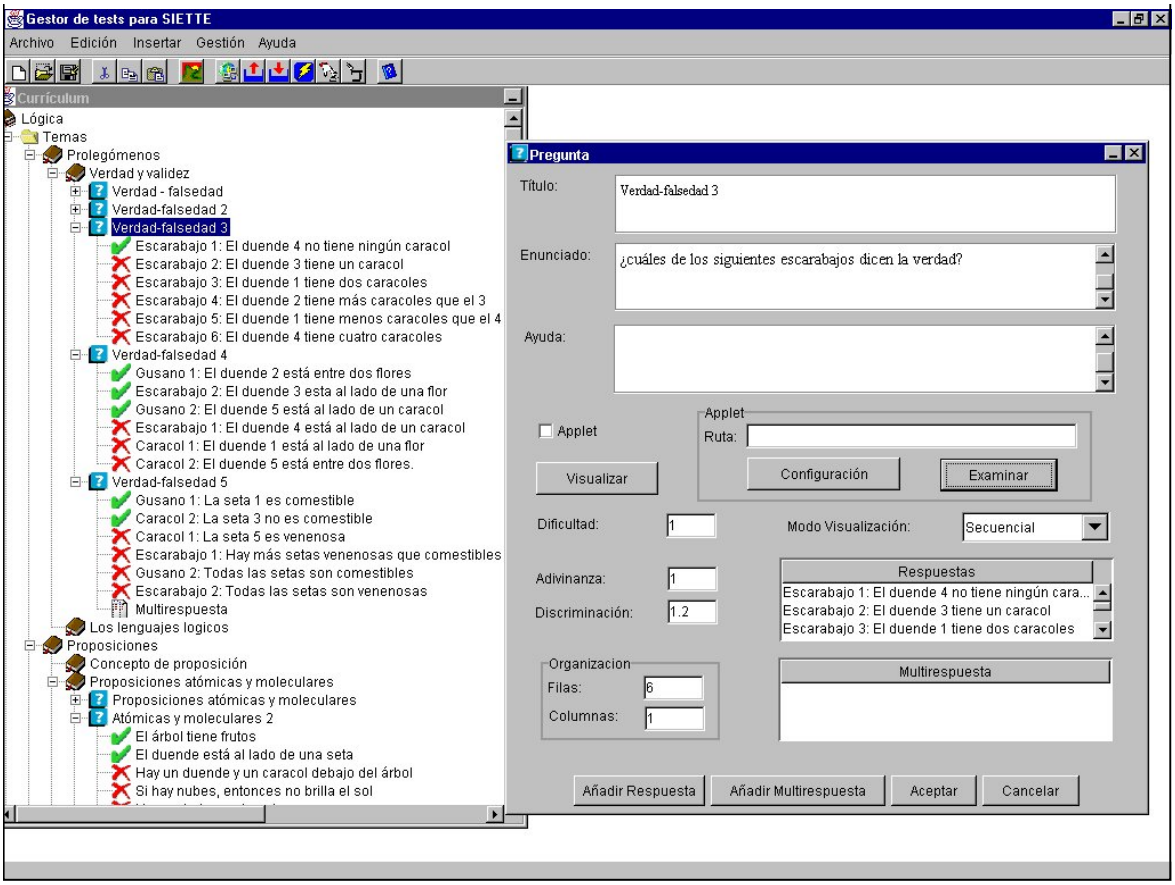

Figura 5. *Interfaz autónoma para edición de tests, temas e ítems.*

Tras especificar los temas, el profesor debe añadir las cuestiones que compondrán el test. En el editor a través de la web tanto el enunciado de las cuestiones como sus respuestas tenían que ser directamente insertados en código HTML. Este nuevo editor ofrece un conjunto de facilidades adicionales de edición del texto, ofreciendo ciertas ventajas al insertar código HTML, objetos multimedia y applets, ya que dispone de un entorno mas elaborado. En concreto, permite configurar fácilmente los parámetros necesarios para un conjunto de applets predefinidos que forman la biblioteca de ítems. (Véase el epígrafe 7)

Una vez que el profesor termina de modificar un test o cualquiera de los componentes de una asignatura, se generará un archivo en formato XML para mantener toda la información. Posteriormente, cuando desee transferir los datos a SIETTE, el editor enviará este archivo en formato XML que se traducirá en modificaciones en la base de conocimientos, mediante el módulo de validación.

En un futuro pueden desarrollarse herramientas de edición específicas para dominios concretos, que sean capaces de gestionar sus propias bibliotecas de ítems, a fin de facilitar la creación del banco de ítems complejos por parte de profesores no programadores.

El módulo de validación y activación de test.

El módulo de validación y actualización de tests es un proceso *fuera de línea* que se ejecuta en el servidor donde esté ubicado el motor de base de datos cada cierto tiempo. Sólo de este modo las modificaciones o creaciones de nuevas especificaciones de tests, realizadas por los profesores, se harán visibles en el generador de tests, manteniendo la coherencia en el caso en que se usen varios editores.

Este módulo comprueba que la definición de las preguntas y los tests es coherente desde el punto de vista computacional, por ejemplo si existe un número suficiente de preguntas de cada tema, si se han incluido suficientes respuestas alternativas para cada pregunta, etc.

Por otra parte la separación entre la base de conocimientos empleada en la generación de test y la base de conocimientos que crea el profesor permite separar las tareas de edición del test de la realización, eliminando los posibles problemas de inconsistencia en el caso de edición y realización simultánea. Un test en ejecución no se ve alterado por la inclusión, modificación o eliminación de ítems.

A continuación se explican detalladamente los mecanismos empleados para realizar la evaluación en SIETTE introduciendo los fundamentos teóricos de la teoría de respuesta al ítem y de la teoría de test adaptativos, indicando las modificaciones realizadas.

3. LA TEORIA DE RESPUESTA AL ITEM

La Teoría Clásica del Test, (TCT), empleada en psicometría fue desarrollada en los años 1920-40. Las deficiencias de esta teoría propiciaron la investigación de modelos alternativos. Una de las alternativas más relevantes fué la Teoría de la Respuesta al Item, (TRI), (Lord & Novick, 1968) inicialmente denominada teoría del rasgo latente. La teoría de la respuesta al ítem se basa en la hipótesis de que la respuesta dada a cada uno de los ítems (cuestiones) del test, depende probabilísticamente de un cierto rasgo que puede medirse mediante un valor numérico fijo, aunque desconocido. Observando las respuestas a cada ítem y conociendo para cada uno de ellos la probabilidad de acierto condicionada al valor del rasgo, puede estimarse un valor probable para el rasgo no directamente observable.

En el campo de la psicometría el rasgo latente θ puede ser cualquier carácter psicológico. En el caso que nos ocupa este rasgo será el *nivel de conocimiento* sobre la materia objeto del test. La definición de este nivel de conocimiento es elusiva. Si bien es una variable normalmente utilizada en la evaluación tradicional de los alumnos no es fácil dar una definición precisa de su significado. Por el momento admítase como hipótesis que el nivel de conocimiento de un alumno sobre una materia puede ser medido mediante una variable real θ que toma valores entre (-∞, +∞). Admítase también que este valor es constante a lo largo de la realización del test, y que las únicas posibles respuestas a una cuestión pueden ser evaluadas como correcto o incorrecto, sin valores intermedios. Volveremos sobre este tema más adelante.

Para poder aplicar la teoría deben suponerse también conocidas para cada ítem del test las probabilidades condicionadas de que una persona con un nivel de conocimiento dado conteste de forma satisfactoria. Esta probabilidad vendrá dada por una función del intervalo (-∞, +∞) en el intervalo real [0,1], y se conoce con el nombre de Curva Característica del Item, CCI. Se exige a estas curvas que sean monótonas crecientes, es decir, que a mayor nivel de conocimiento del alumno, mayor sea la probabilidad de contestar correctamente a la cuestión. Se impone además la condición de independencia entre los ítems, es decir, que la respuesta a una cuestión no esté condicionada por la respuesta que anteriormente se dió a otra cuestión del test.

Los primeros modelos empleados se llamaron modelos normales, puesto que las curvas características empleadas derivaban de la función de distribución normal (Lord & Novick, 1968):

$$P_i(\theta) = P(\mathrm{u_i} = 1 \mid \theta) = \mathrm{c_i} + (1 - \mathrm{c_i}) \frac{1}{a_i \sqrt{2\pi}} \int_{-\infty}^{\theta} e^{-\frac{(\mathrm{x}-\mathrm{b_i})^2}{2 a_i^2}} dx \qquad (1a)$$

Posteriormente se popularizaron otros modelos basados en la función logística. El más conocido es el modelo de un parámetro de Rash (Rash, 1960), y los modelos de dos y tres parámetros (Birnbaum, 1968). En el caso del modelo de tres parámetros las curvas características de los ítems vienen dadas por la fórmula:

$$P_i(\theta) = P(\mathrm{u_i} = 1 \mid \theta) = \mathrm{c_i} + (1 - \mathrm{c_i}) \frac{1}{1 + \mathrm{e}^{-1.7 \mathrm{a_i} (\theta - \mathrm{b_i})}} \qquad (1b)$$

en donde $P(u_i=1 \mid \theta)$ representa la probabilidad de que un alumno cuyo nivel de conocimiento sea θ dé una respuesta acertada al ítem i; en adelante se empleará $P_i(\theta)$ como notación mas compacta para esta función. En principio se asume que todos los ítems son dicotómicos, es decir, que sólo tienen dos posibles respuestas: verdadero o falso. Los tres parámetros de la curva característica del ítem i son: a_i, que se conoce como *factor de discriminación*; b_i también llamado *dificultad* de la pregunta; y c_i, denominado *factor de adivinanza*. La figura 6 muestra la interpretación gráfica del significado de estos parámetros, y el efecto de su aumento o decremento.

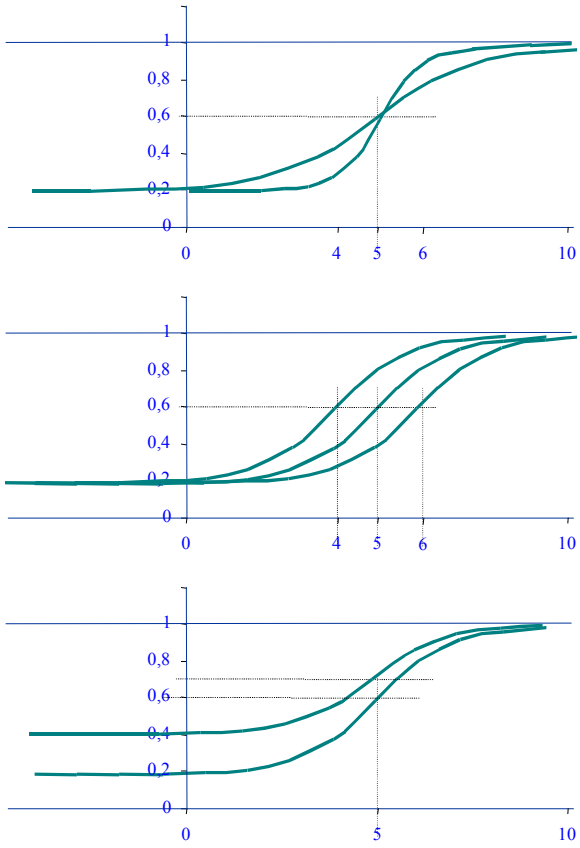

Figura 6. *Parámetros de las Curvas Características del Ítem, como función logística*

El *factor de adivinanza* resulta ser la probabilidad de que un alumno sin ningún conocimiento de la materia responda correctamente a la pregunta. La *dificultad* corresponde al valor del nivel del conocimiento para el cual es equiprobable acertar o errar la respuesta, descontando los efectos del azar. El *factor de discriminación* es proporcional a la pendiente de la curva: cuanto mayor es este factor mayor será la probabilidad de que alumnos con un nivel de conocimiento mayor que la dificultad de la pregunta acierten la respuesta, y que alumnos con menor conocimiento que la dificultad de la pregunta la fallen. Los valores de estos parámetros para cada curva se obtienen de acuerdo a los resultados obtenidos con una población base.

Las respuestas del alumno a cada uno de los ítems del test pueden usarse para estimar el valor del rasgo latente θ. Hay varios métodos para conseguir esto:

- *Método de la máxima verosimilitud* (Lord, 1980), que consiste en obtener el valor de θ que maximiza la función de verosimilitud:

$$L(\mathbf{u}\,|\,\theta) = L(u_1 \ldots u_n\,|\,\theta) = \prod_{i=1}^{n} \mathrm{P_i}(\theta)^{\mathrm{u_i}} (1 - P_i(\theta))^{(1-u_i)}\,, \qquad \text{(2a)}$$

  donde $u = (u_1, \ldots, u_n)$ es el vector de respuestas del alumno, es decir, para cada $i = 1,\ldots,n$, u_i es *1* si la respuesta al ítem i-ésimo es correcta, y *0* en otro caso. Este máximo se obtiene como solución de la ecuación resultante de la derivada de la función de verosimilitud igualada a cero.

$$\left(\frac{\partial(L(\mathbf{u}\,|\,\theta))}{\partial\theta}\right)_{\theta=\hat{\theta}} = 0 \qquad \text{(2b)}$$

- *Método bayesiano*, (Owen 1975), consiste en calcular aplicando la regla de Bayes la distribución de probabilidad del nivel de conocimiento del alumno con posterioridad a la respuesta a cada una de las cuestiones del test.

$$P(\theta \mid \mathbf{u}) = P(\theta \mid u_1 \ldots u_n) = \frac{\prod_{i=1}^{n} \mathrm{P_i}(\theta)^{\mathrm{u_i}} (1 - P_i(\theta))^{(1-u_i)}}{\prod_{i=1}^{n} \mathrm{P}(u_i)^{\mathrm{u_i}} (1 - \mathrm{P}(u_i))^{(1-u_i)}} \mathrm{P}(\theta) \qquad (3)$$

Finalmente se toma como estimador el valor máximo de esta distribución, que como se ve resulta ser proporcional al producto de la función de verosimilitud *L(θ/**u**)* por la función de densidad a priori *P(θ)*.

$$P(\theta/\mathbf{u}) \propto L(\theta / \mathbf{u})\, P(\theta). \qquad (4)$$

Por lo que se obtienen resultados congruentes por ambos métodos.

- *Método basado en redes neuronales*. (Benitez,2000) Alternativamente a los métodos clásicos se ha propuesto la utilización de redes neuronales competitivas y redes neuronales de Kohonen para realizar la clasificación, ya que teóricamente, se ha demostrado, de forma genérica, que los resultados obtenidos con redes neuronales competitivas asemejan a los que se obtienen con la clasificación bayesiana (Funahashi, 1998). La principal ventaja de esta técnica es la aplicación directa de los mecanismos de aprendizaje para la estimación de parámetros. El uso de éste y otros tipos de estructuras de redes y del comportamiento de los mecanismos de aprendizaje es actualmente otra línea de investigación abierta en SIETTE.

Aunque estas ecuaciones (3) y (4) pueden ser resueltas algebraicamente, el cálculo resulta complicado y computacionalmente costoso. Por otra parte para la mayoría de los sistemas tutores inteligentes no es necesaria una precisión muy grande en la evaluación. En general basta con un conjunto de valores discretos. En SIETTE se asume que el rasgo latente θ puede tomar solamente K valores discretos entre *0* y *K-1*. Se considera que las curvas características de los ítems vienen dadas por vectores de K componentes

$$\mathbf{P_i} = \overline{P(u_i \mid \boldsymbol{\theta})} = \left(\Pr(u_i = 1 \mid \theta = 0),\quad \Pr(u_i = 1 \mid \theta = 1),\quad \ldots,\quad \Pr(u_i = 1 \mid \theta = \mathrm{K}\text{-}1)\right) \qquad (5a)$$

siendo el vector complementario a 1:

$$\mathbf{Q_i} = \overline{Q(u_i \mid \boldsymbol{\theta})} = \left(1 - \Pr(u_i = 1 \mid \theta = 0),\quad 1 - \Pr(u_i = 1 \mid \theta = 1),\quad \ldots,\quad 1 - \Pr(u_i = 1 \mid \theta = \mathrm{K}\text{-}1)\right) \qquad (5b)$$

y sea el vector que define la probabilidad a priori:

$$\mathbf{P} = \overline{P(\boldsymbol{\theta})} = \left(\Pr(\theta = 0),\quad \Pr(\theta = 1),\quad \ldots,\quad \Pr(\theta = \mathrm{K}\text{-}1)\right) \qquad (5c)$$

La estimación del valor latente se obtiene mediante el método bayesiano, que queda simplificado en este caso al producto escalar de los vectores característicos de los ítems por el vector de densidad a priori, normalizado de forma que la suma de probabilidades sea *1*. Por lo que la fórmula (3) puede expresarse como:

$$\overline{P(\boldsymbol{\theta}|\mathbf{u})} \propto \mathbf{P} \prod_{i=1}^{n} \mathbf{P_i}^{u_i} \mathbf{Q}_i^{1-u_i} \qquad (5)$$

Ejemplo 1 Veamos un ejemplo para la clasificación del conocimiento de un alumno en cuatro niveles de conocimiento *K=4*, utilizando un test de cinco cuestiones *N=5*. Sean los valores de las curvas características de los 5 ítems dadas en la siguiente tabla:

	θ=0	**θ=1**	**θ=2**	**θ=3**
$\mathbf{P_1}$	0.1	0,3	0,7	0,9
$\mathbf{P_2}$	0,5	0,6	0.9	1,0
$\mathbf{P_3}$	0,3	0,6	0,8	0,9
$\mathbf{P_4}$	0,3	0,4	0,7	0,9
$\mathbf{P_5}$	0,1	0,2	0,3	0,9

Supongamos que a priori (antes de realizar el test) estimamos equiprobable cualquier valor de θ y que el vector de respuestas del alumno *A* es ($u_1=1$, $u_2=1$, $u_3=0$, $u_4=1$, $u_5=0$). Aplicando 5 veces la regla de Bayes se obtiene:

$$\overline{P(\theta \mid u_1 \ldots u_5)} \propto \begin{pmatrix} 0{,}1 \\ 0{,}3 \\ 0{,}7 \\ 0{,}9 \end{pmatrix} \begin{pmatrix} 0{,}5 \\ 0{,}6 \\ 0{,}9 \\ 1{,}0 \end{pmatrix} \begin{pmatrix} 0{,}7 \\ 0{,}4 \\ 0{,}2 \\ 0{,}1 \end{pmatrix} \begin{pmatrix} 0{,}3 \\ 0{,}4 \\ 0{,}7 \\ 0{,}9 \end{pmatrix} \begin{pmatrix} 0{,}9 \\ 0{,}8 \\ 0{,}7 \\ 0{,}1 \end{pmatrix} (0{,}25 \quad 0{,}25 \quad 0{,}25 \quad 0{,}25) \propto (0{,}092 \quad 0{,}225 \quad 0{,}603 \quad 0{,}079)$$

Es decir, que con una probabilidad de *0,603* el nivel de conocimiento del alumno *A* es $\theta=2$

El uso de distribuciones discretas en vez de funciones continuas, además de ser más sencillo de implementar tiene ventajas adicionales como veremos más adelante. En principio, las curvas características de los ítems no requieren de parámetros para su definición, y por tanto la utilización de una función de distribución normal o de una función logística como la propuesta en (1a) o (1b) no es una elección que deba hacerse a priori, sino que puede estimarse sin restricciones a partir de las observaciones como veremos más adelante.

Otra característica muy interesante de la aproximación discreta es que el número de valores *K* que puede tomar el rasgo latente θ, en este caso el nivel de conocimiento, no necesita prefijarse totalmente al definir el test, sino que puede elegirse en el momento de iniciar cada uno de los tests. Evidentemente, el número de valores de cualquier test no puede ser superior a los *K* que se almacenan para definir las curvas características de cada ítem, pero dado que se trata de un producto escalar, establecer un valor K' submúltiplo de *K* equivale a agrupar K/K' valores consecutivos de las curvas características, sumando los valores, es decir, si las curvas características están definidas como vectores de dimensión *K*: $\overline{P(\theta)} = (p_0, p_1, \ldots p_{k-1})$.y se desea evaluar a un alumno mediante K' niveles, las curvas características que deben emplear serán: $\overline{P'(\theta)} = (p'_0, p'_1, \ldots p'_{k'-1})$, en donde

$$p'_j = K'/K \times (p_{j \times K/K'} + p_{j \times K/K'+1}, \ldots + p_{j \times K/K'+K/K'-1}) \qquad (6)$$

Ejemplo 2 Siguiendo con el ejemplo anterior, si se decide clasificar al alumno solamente mediante dos niveles de conocimiento $K'=2$, y suponiendo el mismo vector de respuestas a las preguntas del ejemplo anterior, bastaría con realizar las operaciones:

$$P(\theta \mid u_1 \ldots u_5) \propto \begin{pmatrix} 0{,}20 \\ 0{,}80 \end{pmatrix} \begin{pmatrix} 0{,}55 \\ 0{,}95 \end{pmatrix} \begin{pmatrix} 0{,}65 \\ 0{,}15 \end{pmatrix} \begin{pmatrix} 0{,}35 \\ 0{,}80 \end{pmatrix} \begin{pmatrix} 0{,}85 \\ 0{,}40 \end{pmatrix} (0{,}5 \quad 0{,}5) \propto (0{,}317 \quad 0{,}682)$$

Como puede verse esto ofrece una ventaja computacional considerable frente al método basado en distribuciones reales, en las que la complejidad computacional es sólo función del número de ítems del test.

Aunque no son estrictamente necesarios, SIETTE emplea tres parámetros por analogía con el modelo mas extendido de la teoría de respuesta al ítem, para la definición heurística de las curvas características de los ítems. Dado que SIETTE emplea solamente *K* niveles de conocimiento las curvas características se aproximan mediante los valores de la curva real en puntos equidistantes del intervalo $[-(K-1)/2., (K-1)/2)]$. SIETTE puede utilizar tanto aproximaciones basadas la función de distribución normal (1a) como en la función logística de la ecuación (1b). Por ejemplo, en este segundo caso el vector característico de un ítem en SIETTE $P(\theta) = (p_0, p_1, \ldots p_{K-1})$, puede definirse inicialmente mediante la función:

$$p_k = p_k(u=1 \mid \theta = k) = c + (1-c)\frac{1}{1+e^{-1.7a\left(\left(k-\frac{K-1}{2}\right)-b\right)}} \qquad (7)$$

Puesto que SIETTE solo emplea *K* valores de conocimiento se exige que el parámetro de dificultad de todas las curvas tome también valores entre *0* y *K-1*.

También está implementado, aunque por el momento no se usa, un cuarto factor, al que se ha denominado f*actor de distracción*, y que representa la probabilidad de que un alumno con conocimiento máximo falle la pregunta como consecuencia de una distracción, y no de su falta de conocimiento. Este parámetro se trata análogamente al factor de adivinanza en las fórmulas (1a) y (1b).

Como veremos más adelante, SIETTE es capaz de modificar los vectores característicos de los ítems a partir de los datos obtenidos del uso del sistema. Esto modifica la forma de las curvas características, que ya no pertenecerán a la familia normal o logística seleccionada inicialmente. Si se desea, se puede configurar SIETTE de manera que mantenga siempre vectores correspondientes a curvas normales o logísticas de la familia de tres parámetros, como se hace en la teoría clásica de respuesta al ítem. Este proceso, inverso al anterior, se realiza mediante ajuste por mínimos cuadrados, entre el vector característico en uso, y los vectores resultantes del proceso de discretización expuesto anteriormente.

4. LA TEORIA DE TEST ADAPTATIVOS

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 tiene la ventaja de que todos los examinandos están en igualdad de condiciones. Como contrapartida, este tipo de tests debe 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 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 la creación de tests más flexibles que aliviarían en parte estos problemas. Lord (Lord, 1970) trabajó en 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, 90), 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 Test Adaptativos Informatizados, TAI, o tests administrados 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 test adaptativo informatizado es un algoritmo iterativo que comienza con una estimación inicial del nivel de conocimiento del alumno y que tiene los siguientes pasos (véase figura 7):

1. Todas las preguntas que no se han administrado todavía, son examinadas para determinar cual será la mejor para ser propuesta la siguiente, según la estimación hecha hasta el momento del nivel de conocimiento 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.
4. Los pasos desde el 1 al 3 son repetidos hasta que se cumpla alguno de los criterios de terminación definidos.

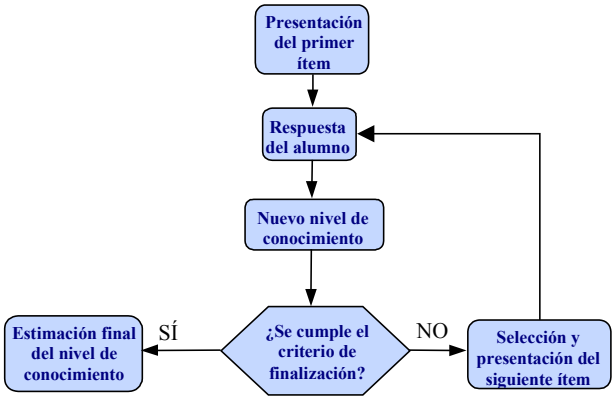

Figura 7. *Diagrama de flujo de un test adaptativo. Adaptado de (Olea & Ponsoda, 1996).*

Los elementos básicos de un test adaptativo informatizado son:

- *Banco de preguntas*. Constituye uno de los elementos fundamentales para la creación de un test adaptativo. El banco de preguntas debe contener ítems en suficiente número, variedad y niveles de dificultad. (Flaugher, 1990).
- *Modelo de respuesta* asociado a cada pregunta. Es decir, el modelo que predice como el sujeto responderá al ítem según su nivel de conocimiento.
- *Criterio 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 mejora la precisión en la estimación del nivel de conocimiento y reduce la longitud del test.
- *Criterio de terminación*. Para decidir cuando 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.
- *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 de estimación, como considerar el nivel medio de los sujetos que han realizado el test previamente, o crear un perfil del sujeto y usar el nivel medio de los alumnos con un perfil similar (Thissen, D. & Mislevy R.J., 1990). La estimación inicial tiene mayor influencia cuanto menor es el número de preguntas del test.

Es muy habitual utilizar la Teoría de la Respuesta al Ítem como modelo de respuesta de los Test Adaptativos Informatizados. Ambas teorías se complementan y por tanto cuando se habla de *tests adaptativos informatizados*, en general se presupone que el modelo de respuesta que subyace es el modelo de la *teoría de respuesta al ítem*.

En cuanto a los métodos de selección los más comunes son:

- *Método de máxima información* (Weiss, 1982) que consiste en seleccionar el ítem que haga máxima la información del ítem para el nivel de conocimiento estimado hasta el momento, definiendo una cierta f*unción de información* para cada ítem en función del conocimiento estimado del alumno en cada instante.
- *Métodos Bayesianos*, como el propuesto por (Owen, 1975), que selecciona la pregunta que hace mínima la esperanza matemática de la varianza del nivel de conocimiento del alumno a posteriori. En el caso de curvas características definidas sobre la recta real mediante la función logística de tres parámetros, se demuestra que la mínima varianza se alcanza a igualdad de otros parámetros con el ítem cuyo parámetro de dificultad es más próximo a la estimación del nivel de conocimiento del alumno en ese momento; o con aquel con un mayor factor de discriminación (a igualdad de los otros parámetros), o al de menor factor de adivinanza (a igualdad de los demás parámetros).

Los tests adaptativos tienen importantes ventajas frente a los tests no adaptativos, principalmente un decremento significativo en el número de ítems para alcanzar estimaciones con una precisión dada, o bien estimaciones más precisas del nivel de conocimiento del alumno que tests no adaptativos de igual número de preguntas. Por otra parte, ofrece una mejora en la motivación de los alumnos, al plantear preguntas adecuadas a su nivel de conocimiento.

SIETTE es una implementación el algoritmo de tests adaptativos informatizados basado en el modelo de respuesta de la teoría de respuesta al ítem discreta expuesta en el apartado anterior. El *criterio de selección* de preguntas es configurable para cada test a elección entre los siguientes:

- *Bayesiano*, dada la distribución de conocimiento del alumno, para todas las preguntas no planteadas se elige aquella que hace mínima la suma de varianzas a posteriori de las distribuciones resultantes del acierto o fallo del ítem.
- *Basado en la dificultad*, es un método basado en máxima información, pero considerando como tal al ítem cuya dificultad es más próxima en valor absoluto al valor más probable del nivel de conocimiento del alumno, (en caso de igualdad de distancia, se elige aleatoriamente entre los candidatos)
- *Aleatorio*, esta opción permite desactivar el mecanismo de adaptación.

Ejemplo 3. Sean los vectores característicos de las 5 cuestiones del ejemplo 1. Supuesto que en un instante intermedio del test el nivel de conocimientos del alumno se estima mediante el vector (0,1; 0,2; 0,6; 0,1); para aplicar el método de selección bayesiano, se calculan los vectores resultantes a posteriori suponiendo respuestas correctas e incorrectas a cada pregunta, se calcula la esperanza matemática de la varianzas de estos vectores para cada pregunta y se escoge la menor, que en este caso resulta ser el ítem 5.

	u=1 (acierto)				media	var.	u=0 (fallo)				media	var.	E[var]
P_1	(0,017	0,103	0,724	0,155)	2,017	0,327	(0,214	0,333	0,429	0,024)	1,262	1,240	0,784
P_2	(0,062	0,148	0,667	0,123)	1,852	0,497	(0,263	0,421	0,316	0,001)	1,053	1,215	0,856
P_3	(0,042	0,167	0,667	0,125)	1,875	0,443	(0,250	0,286	0,429	0,036)	1,250	1,150	0,796
P_4	(0,048	0,129	0,677	0,145)	1,919	0,461	(0,184	0,316	0,474	0,026)	1,342	0,979	0,720
P_5	(0,031	0,125	0,563	0,281)	2,094	0,522	(0,132	0,235	0,618	0,015)	1,515	0,879	0,701

El *criterio de finalización* es igualmente configurable para cada test, seleccionando una combinación de entre los siguientes:

- Valor más probable de la distribución de conocimiento estimado superior a un cierto umbral.
- Varianza de la distribución de conocimiento estimado menor que un valor dado.
- Número mínimo y máximo de preguntas, para garantizar que en cualquier caso el test termina, y que se han realizado un mínimo de preguntas que amortiguan el efecto de la estimación inicial del conocimiento del alumno.

SIETTE asume por defecto que la distribución de conocimiento a priori del alumno antes de efectuar el test es uniforme, lo que equivale a suponer que todos los niveles de conocimiento son inicialmente equiprobables para un alumno cuyo historial se desconoce. Otras opciones son también posibles, como considerar distribuciones normales centradas, o funciones de densidad obtenidas a partir de los datos de la población de alumnos que han realizado el mismo test hasta el momento. SIETTE guarda información sobre la distribución de conocimiento de un alumno concreto y también puede usarla como punto de partida en sucesivas evaluaciones mediante el mismo test.

5. ESTUDIO SIMULADO DEL COMPORTAMIENTO TEÓRICO.

Paralelamente al sistema SIETTE descrito en el apartado 2, se ha desarrollado un programa que simula la realización de tests por parte de hipotéticos alumnos con características conocidas. El objetivo es evaluar y comparar el comportamiento de los algoritmos empleados en la generación de un test, así como analizar empíricamente la influencia de los distintos parámetros que intervienen en él, y estudiar en un entorno controlado nuevos algoritmos de aprendizaje que se describirán en el siguiente epígrafe, o mecanismos para la identificación de ítems anómalos.

Existen numerosos factores que influyen en los tests, tanto en las estimaciones de conocimiento de los alumnos, como en las curvas características de los ítems, o en las características globales del test. Los principales factores que se pueden configurar en el simulador son:

- Número de niveles de conocimiento.
- Tamaño de la base de preguntas del test.
- Criterio de selección de preguntas: aleatorio, basado en la dificultad o bayesiano.
- Criterio de finalización: fijo o variable en función de la precisión.
- Método de evaluación: modal, bayesiano o porcentual.
- Distribución del conocimiento real de los alumnos: homogénea o binomial.
- Número de alumnos (= número de test a realizar).
- Conjunto inicial de preguntas.
- Dificultad (real o estimada) de los ítems.
- Factor de discriminación (real o estimado).
- Factor de adivinanza (real o estimado).
- Factor de distracción (real o estimado).
- Modo de aprendizaje: incremental, por lotes, o no incremental

El simulador de tests, se compone de cuatro fases: (a) generación del conjunto de ítems, (b) simulación del test, (c) aprendizaje de las distribuciones, y (d) análisis de resultados.

En el simulador, cada ítem tiene asociadas tres curvas o vectores característicos que se utilizan a lo largo de la simulación:

- *Curva característica real del ítem* que representa el comportamiento real del ítem. Un alumno de conocimiento real *k*, responderá a un determinado ítem de acuerdo con esta curva.

- *Curva característica estimada del ítem* que representa la aproximación que el sistema utiliza en cada momento para aplicar el modelo de respuesta o método de evaluación.
- *Curva característica aprendida del ítem* que se va construyendo a lo largo de la simulación. Representa la modificación sobre la curva estimada obtenida mediante el mecanismo de aprendizaje.

El simulador es capaz de trabajar con bancos de ítems predefinidos, cuyas curvas características reales y estimadas se introducen mediante un fichero. También es posible generar bancos de ítems aleatorios según diversas distribuciones de los parámetros reales y estimados o utilizar bancos de ítems patrones que cumplan determinadas condiciones, entre estos:

- *Conjuntos de ítems correctos***:** en los que las curvas características estimadas coinciden con las reales. A su vez se contemplan dos variantes:
  - *Homogéneo*: si existe el mismo número de ítems para cada nivel de dificultad.
  - *Aleatorio*: en la que los parámetros de las preguntas siguen una distribución normal.
- *Conjuntos de ítems equilibrados***:** en los que los parámetros de dificultad reales corresponden a la media de los parámetros de dificultad estimada para cada nivel de conocimiento *k*. Es decir los parámetros reales y estimados cumplen la siguiente propiedad:

$$\forall k \in [0, K-1] \quad \forall i \in I \quad d_r(i) = \frac{\sum_{d_r(j)=k} d_e(j)}{N_k}$$

  en donde K es el número de componentes del nivel de conocimiento, I es el conjunto de ítems, d_r(i) es el parámetro de dificultad real del ítem i; $d_e(j)$ es el parámetro de dificultad estimado del ítem j, y N_k es el total de ítems cuya dificultad estimada es k.
- *Conjuntos aleatorios*, en los que los valores de los parámetros reales y estimados de los ítems se generan de forma aleatoria e independiente siguiendo una distribución normal o una distribución uniforme.

Una vez que el sistema ha generado el conjunto de preguntas comienza la simulación propiamente dicha. El sistema simula tantos alumnos como tests se le hayan indicado. Cada alumno es representado por su nivel de conocimiento real (valores entre *0* y *K-1*). Una vez seleccionado un ítem se simula la respuesta del alumno, para lo que el sistema extrae de su *curva característica real* la probabilidad de acierto para el nivel de conocimiento real del alumno y genera un número aleatorio con distribución uniforme entre *0* y *1*. Si es menor o igual que la probabilidad de acierto, la respuesta a la pregunta se considera correcta. A continuación se aplica el modelo de respuesta elegido utilizando las curvas características estimadas. En caso necesario el criterio de selección y finalización emplean igualmente las curvas estimadas. Finalmente se hace un pequeño análisis estadístico de los resultados: número de preguntas planteadas, porcentaje de alumnos correctamente clasificados, distancia entre las curvas aprendidas y las reales, etc.

Por ejemplo, se han realizado pruebas con la ayuda del simulador para comprobar y cuantificar los resultados teóricos que se obtienen usando los distintos criterios de selección de ítems usando un banco de preguntas bien calibradas y suponiendo una distribución uniforme de alumnos. En resumen las conclusiones son que el uso del criterio *Bayesiano* o *Basado en la Dificultad* reduce a más de la mitad el número de preguntas necesario manteniendo el mismo nivel de precisión que utilizando el criterio *Aleatorio*. El método *Bayesiano* es ligeramente mejor que el método *Basado en la Dificultad*, aunque este último es considerablemente más eficiente desde el punto de vista computacional. La siguiente tabla muestra los resultados obtenidos en el simulador usando una muestra de 1000 alumnos de nivel de conocimiento distribuido unifórmente, usando un banco de 100 preguntas con dificultades distribuidas uniformemente, en las que el factor de adivinanza es *0* y el factor de discriminación es *1,2*. La clasificación se ha realizado para un factor de confianza del 90%. (Conejo,00)

	Criterio de selección Aleatorio		Criterio de selección Bayesiano		Criterio de Selección Basado en Dificultad	
Número de clases K	*% de alumnos calsificados correctamente*	*Numero medio de preguntas planteadas*	*% de alumnos calsificados correctamente*	*Numero medio de preguntas planteadas*	*% de alumnos calsificados correctamente*	*Numero medio de preguntas planteadas*
3	95.82	3.59	96.06	3.58	95.62	3.58
5	92.76	10.38	93.31	6.87	94.67	7.37
7	92.85	18.16	92.75	8.70	94.43	9.03
9	92.93	26.39	92.53	9.85	94.23	10.14
11	92.92	34.54	92.10	10.71	94.14	11.02

Table 1. *Precisión y número de preguntas según el criterio de selección*

6. ESTIMACIÓN DE LAS CURVAS CARACTERÍSTICAS DE LOS ÍTEMS.

Por lo que respecta a la estimación de los parámetros de los ítems, dentro de la teoría de respuesta al ítem, los métodos clásicos son de dos tipos principalmente (Santiesteban, 1990):

- *Métodos condicionales*: Consisten en construir la función de verosimilitud considerando como incógnitas tanto a los parámetros de las preguntas como los rasgos latentes de un conjunto de individuos que realizan el test, es decir se establece una condición entre los rasgos latentes y los parámetros de los ítems. Por ejemplo, en el caso de utilizar el modelo de tres parámetros dado por la ecuación (1b), supuesta una población de *m* individuos que realiza cada uno de ellos un test con *n* preguntas, la función de verosimilitud es el producto de las funciones de verosimilitud de cada uno de los tests individuales (2a). Si suponemos desconocidos tanto los *3×n* parámetros de las curvas características de los ítems $a_1,b_1,c_1 \ldots a_n,b_n,c_n$ como los *m* valores del rasgo latente de cada uno de los alumnos $\theta_1,\ldots,\theta_m$, la función de verosimilitud resulta ser:

$$L(u_{ij}/\theta_j,a_i,b_i,c_i) = L(u_{11},\ldots,u_{n1},u_{12},\ldots,u_{n2},\ldots u_{1m},\ldots,u_{nm}/\theta_1,\ldots,\theta_m,a_1,b_1,c_1\ldots,a_n,b_n,c_n) =$$
$$= \prod_{j=1}^{m}\prod_{i=1}^{n} \mathrm{P_i}(\theta_j)^{\mathrm{u_{ij}}}(1-P_i(\theta_j))^{(1-u_{ij})} \qquad (8)$$

  El máximo de la función de verosimilitud se obtiene derivando parcialmente con respecto a cada uno de las *3×n+m* incógnitas obteniendo un sistema de *3×n+m* ecuaciones:

$$\left(\frac{\partial \ln L(\mathrm{u}_{ij}\mid\theta_j,a_i,b_i,c_i)}{\partial a_i}\right)=0 \quad \left(\frac{\partial \ln L(\mathrm{u}_{ij}\mid\theta_j,a_i,b_i,c_i)}{\partial b_i}\right)=0$$
$$\left(\frac{\partial \ln L(\mathrm{u}_{ij}\mid\theta_j,a_i,b_i,c_i)}{\partial c_i}\right)=0 \quad \left(\frac{\partial \ln L(\mathrm{u}_{ij}\mid\theta_j,a_i,b_i,c_i)}{\partial \theta_j}\right)=0 \qquad (9)$$

  en donde como es habitual se han tomado logaritmos, dado que el máximo de una función real no negativa se alcanza en el mismo punto que el máximo del logaritmo de esa función y el cálculo es así mas sencillo.

- *Métodos incondicionales*, estos métodos se basan en asumir un conocimiento a priori sobre la distribución de la población de individuos que realiza el test, *h(θ)*, por ejemplo, suele asumirse que el conjunto de valores del rasgo latente sigue una distribución normal en el conjunto de la población. Esta distribución debe integrarse en la función de verosimilitud, por lo que ésta solo dependerá de los parámetros de los ítems:

$$L(u_{ij}/a_i,b_i,c_i) = L(u_{11},\ldots,u_{n1},u_{12},\ldots,u_{n2},\ldots u_{1m},\ldots,u_{nm}/a_1,b_1,c_1\ldots,a_n,b_n,c_n) =$$
$$= \prod_{j=1}^{m} \int \mathrm{P_i}(\theta)^{\mathrm{u_{ij}}}(1-P_i(\theta))^{(1-u_{ij})}\, h(\theta)\, d\theta \qquad (10)$$

  al igual que en el caso anterior para hallar el máximo de la función de verosimilitud se deriva parcialmente el logaritmo de esta función con respecto a cada uno de los parámetros, obteniendo un sistema de *3×n* ecuaciones con *3×n* incógnitas en el caso del modelo de respuesta dado por (2a)

Para resolver estas ecuaciones se usan métodos numéricos por aproximaciones sucesivas, aunque resultan complicados y computacionalmente costosos. Cada uno de estos métodos tiene sus inconvenientes. En los métodos condicionales aumenta el número de incógnitas cuando aumenta el tamaño de la muestra. En los métodos incondicionales se hace necesario estimar a priori la forma de la densidad *h(θ),* es decir cuantos alumnos de cada nivel de conocimientos componen la población.

En SIETTE se aborda el problema de forma diferente. Pero antes de abordar el tema de la calibración debemos reflexionar sobre el objetivo final, es decir la estimación de un cierto rasgo latente en una población a partir de las curvas características de los ítems calibradas con los datos de una muestra inicial de control. Este rasgo latente en SIETTE es el *nivel de conocimiento* del alumno sobre una determinada materia. No existe una definición clara del

significado de este valor. La teoría clásica le asigna un valor real, la aproximación de SIETTE es más simple, al utilizar solamente un conjunto finito de clases de valores para el nivel de conocimiento, pero desde el punto de vista conceptual el problema es el mismo.

Una posible definición para el *nivel de conocimiento* del alumno sería el porcentaje de cuestiones que es capaz de contestar correctamente sobre el total de un *banco de cuestiones* sobre la materia objeto de estudio. En este caso el nivel de conocimiento del alumno se define en función del banco de preguntas. Si suponemos que el banco de preguntas es suficientemente extenso y completo, esta definición resulta operativa. El propósito del test sería estimar este porcentaje utilizando tan solo un número reducido de cuestiones de todas las que componen el banco. Según *teoría clásica de los tests*, si las preguntas se seleccionan de forma aleatoria, la media de aciertos sobre la muestra correspondería al estimador de máxima verosimilitud del nivel de conocimientos. La aplicación de la teoría de los test adaptativos reduce el número de cuestiones necesarias para la obtención de este estimador, a costa de la calibración de las curvas características.

Esta calibración puede realizarse fácilmente a partir de un conjunto inicial de *m* individuos a los que se plantean preguntas de forma aleatoria; utilizando el estimador resultante como el valor real del nivel de conocimiento *k*, y estimando directamente los valores de cada componente del vector característico del ítem, según la formula:

$$p_{ik} = p_{ik}(u_i = 1 \mid \theta = k) = \frac{m(u_i = 1/\theta = k)}{m(\theta = k)} \qquad (11)$$

en donde, $p_{ik}(u_i=1|\theta=k)$, es la componente *k*-ésima del ítem *í*; $m(\theta=k)$ es el total de alumnos que han obtenido una puntuación *k* en el test, y $m(u_i=1|\theta=k)$ es el total de alumnos que han respondido correctamente al ítem *í* y han finalizado el test con una puntuación *k*.

Otra posible definición para el *nivel de conocimiento* del alumno podría hacerse de forma indirecta, desde un punto de vista puramente funcional. Puede suponerse en este caso que es el profesor el que mediante algún conjunto de condiciones define a qué clase pertenece cada uno de sus alumnos a efectos de las acciones instructoras a realizar. Por ejemplo, el profesor puede expresar una condición en la forma: "*Todos los alumnos que contesten correctamente a esta pregunta son del grupo avanzado*". Extrapolando esta idea, el profesor podría definir a priori todos los parámetros de las cuestiones por estimación directa, ya que al fin y al cabo no son más que condiciones estadísticas que se imponen para la pertenencia a tal o cual clase. Sin embargo, los grados de libertad del sistema no son tantos. Una vez fijada la curva característica de una cuestión, hipotéticamente podría obtenerse el nivel de conocimiento de un alumno por repetición de la misma, por lo que los parámetros de una nueva cuestión estarían condicionados a la anterior. Matemáticamente lo que ocurre es que el sistema de ecuaciones en (9) se haría incompatible, al tener mas ecuaciones que incógnitas. Una solución es suponer que los parámetros de las curvas siguen también una cierta distribución, introducir las funciones de densidad en (8) e integrar (Swaminathan y Gilford, 1981).

En este caso, la solución adoptada en SIETTE para la calibración consiste en considerar para las curvas características dos ternas de parámetros. Un conjunto de parámetros se denomina *parámetros estimados*, y son los que proporciona el profesor inicialmente. A los otros tres se les denomina *parámetros reales*. La hipótesis en SIETTE es que el profesor ha calibrado la dificultad de las preguntas de forma no sesgada, es decir, se supone que el profesor ha cometido errores en la asignación del parámetro inicial de dificultad de las preguntas, pero que la distribución de estos errores sigue una normal cuya media es precisamente el valor de dificultad real, o lo que es lo mismo, la dificultad real de un conjunto de cuestiones de un test que se comportan frente a un conjunto de estudiantes como ítems de una misma dificultad, corresponde a la media de las dificultades estimadas asignadas inicialmente por el profesor. A un banco de preguntas que cumple esta condición se le ha denominado *conjunto de ítems equilibrado*. (Véase el epígrafe anterior).

Se han realizado estudios empíricos con la ayuda del simulador descrito en el anteriormente sobre el error que se comete al utilizar los estimadores de los parámetros de las preguntas en lugar de los parámetros reales para un conjunto equilibrado de preguntas. La hipótesis del conjunto equilibrado solo impone condiciones sobre el parámetro de dificultad, pidiendo al profesor una estimación subjetiva del mismo. No se exige al profesor que estime los parámetros de discriminación y de adivinanza. Este último puede en muchos casos estimarse dependiendo del tipo de ítem. Por ejemplo en una pregunta con múltiples opciones mutuamente excluyentes el factor de adivinanza puede estimarse como *1/r*, siendo *r* el número de opciones. Los resultados empíricos muestran que las estimaciones del nivel de conocimiento obtenidas mediante conjuntos equilibrados de cuestiones se aproximan mucho a las que se obtendrían con los valores reales de los parámetros. (Conejo 00). Se ha estudiado además el

efecto de un conjunto equilibrado de cuestiones frente a un conjunto correcto, es decir, sin errores en la estimación. La conclusión es que la fiabilidad del test no se ve comprometida sustancialmente, pero se requiere un mayor número de cuestiones para conjuntos equilibrados, y el algoritmo de adaptación pierde efectividad. También se observan mejores resultados en conjuntos equilibrados de ítems si inicialmente se estiman valores moderadamente bajos para los coeficientes de discriminación, independientemente de cual sea su valor real. (véase tabla 2)

	Conjunto correcto de ítems		Conjunto equilibrado de ítems			
	Criterio de Selección Aleatorio		Criterio de Selección Aleatorio		Criterio de Selección Basado en Dificultad	
Factor de discriminación estimado a_e	*% de alumnos calsificados correctamente*	*Numero medio de preguntas planteadas*	*% de alumnos calsificados correctamente*	*Numero medio de preguntas planteadas*	*% de alumnos calsificados correctamente*	*Numero medio de preguntas planteadas*
0.2	90.4	174.9	55.4	78.2	85.4	186.8
0.5	91.5	35.2	83.1	32.1	82.4	33.3
0.7	91.9	26.3	85.4	25.8	81.1	18.3
1.2	92.8	18.1	83.1	16.0	78.4	8.6
1.7	93.8	15.3	73.7	12.0	71.4	6.1

Table 2. *Resultados obtenidos mediante conjuntos de ítem correctos y equilibrados*

Como corolario, los valores obtenidos como estimadores del nivel de conocimiento del alumno, usando conjuntos equilibrados de ítems, pueden utilizarse del mismo modo que en (11) para el aprendizaje de las componentes de los vectores característicos de los ítems. Se han estudiado tres modos de aprendizaje dependiendo del momento en el que éste se realice a lo largo de la simulación:

- *incremental***:** cada vez que se realiza un test se modifican las curvas características de las preguntas que han intervenido en dicho test;
- *por lotes***:** la modificación de la curva característica se realiza cada *m* tests.
- *no incremental***:** la modificación de las curvas de las preguntas se realiza sólo al final de la muestra;

Se ha comprobado empíricamente, mediante el simulador, que para conjuntos equilibrados de ítems, todos estos mecanismos de aprendizaje convergen hacia la estimación de los parámetros reales, siendo el procedimiento incremental el que converge mas rápidamente. Sin embargo, el número de casos de la muestra necesarios para la estimación es mas elevado que en el caso de aproximaciones mediante el método de la máxima verosimilitud. Esto se debe a que al estimar las componentes del vector característico directamente, existen muchos mas grados de libertad. Si se impone como condición que los vectores característicos sigan perteneciendo a la familia de distribuciones logísticas de tres parámetros, el número de casos necesarios para una buena calibración se reduce sustancialmente, especialmente para vectores con un mayor número de componentes. En este caso tras aplicar la fórmula (11) se aproxima por mínimos cuadrados la curva de cada ítem utilizando todos los casos disponibles. Este procedimiento requiere el modo de aprendizaje por lotes.

La conclusión de todo ello es que si se acepta la hipótesis del conjunto equilibrado, solo se requiere un valor estimativo de la dificultad de las preguntas por parte del profesor. Los factores de adivinanza pueden estimarse de acuerdo al tipo de ítems y los valores del factor de discriminación deben inicialmente estimarse a la baja. Una vez preparado el test puede ser incluso conveniente utilizar un criterio de selección aleatorio en las primeras ejecuciones sesiones, incorporando un mecanismo de aprendizaje incremental tras la finalización de cada prueba, que converge hacia los valores reales de los parámetros. La demostración formal de estas propiedades y la cuantificación de los errores es una de las líneas de investigación actualmente abiertas en relación con SIETTE.

7. TIPOS DE ÍTEMS

SIETTE es capaz de trabajar con diferentes tipos de ítems, y combinarlos en el mismo test utilizando el mismo modelo de respuesta. Los tipos de ítems son:

- ***Ítems dicotómicos*** Hasta el momento siempre que se ha hecho alusión a los ítems, se ha considerado que se trata de cuestiones dicotómicas, cuya respuesta es simplemente verdadero o falso. Desde el punto de vista de las curvas características, los ítems dicotómicos en SIETTE se definen mediante un único vector, tal como se ha expuesto al final del tercer apartado. Este vector *$(p(u=1|\theta=0), p(u=1|\theta=1),... p(u=1|\theta=K-1))$* se utiliza para el cálculo de la probabilidad a posterior en el caso de que la pregunta sea contestada acertadamente, y en el caso de fallo se utiliza el vector complementario, es decir: *$(p(u=0|\theta=0), p(u=0|\theta=1),... p(u=0|\theta=K-1))$* en donde

cada componente p_k se obtiene como $p_k(u=0)= 1- p_k(u=1)$. El criterio de selección es el que se expuesto ampliamente en el epígrafe 4. Un ejemplo de este tipo de ítems se muestra en la siguiente figura:

Figura 8. *Items dicotómicos*

- ***Ítems politómicos*** Son aquellos que tienen más de dos respuestas para un mismo enunciado, entre las cuales el alumno debe seleccionar sólo una. SIETTE dispone las alternativas aleatoriamente cada vez que muestra la pregunta. (Thissen y Steinberg 1997) Un ejemplo de este tipo de ítems se muestra en la siguiente figura:

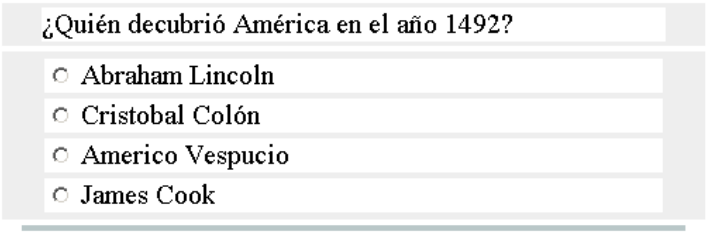

Figura 9. *Items politómicos*

Desde el punto de vista de las curvas características, los ítems politómicos en SIETTE se definen asociando a cada posible respuesta un vector característico de probabilidad condicionada. Las posibles respuestas a un ítem politómico con r respuestas alternativas $a_0 ... a_r$ son $r+1$, una por cada una de las posibles respuestas y otra más que corresponde a la contestación en blanco. Para aplicar el modelo de respuesta es necesario definir las $r+1$ curvas características $(p_0(u=a_j), p_1(u=a_j),... p_{K-1}(u=a_j))$. Para cada componente k, la suma de las componentes k-ésimas de todos los vectores debe ser 1, dado que son todas las opciones posibles. Para el cálculo de la probabilidad a posterior se utiliza el vector correspondiente dependiendo de la respuesta del alumno. La estimación inicial de estas $r+1$ curvas características, se hace en función de la curva característica de la opción correspondiente a la respuesta correcta. El profesor debe indicar cual de las opciones es la respuesta correcta, a_c y estimar mediante el parámetro de dificultad, el vector $(p_0(u=a_c), p_1(u=a_c),... p_{K-1}(u=a_c))$. El resto de las curvas se estiman repartiendo uniformemente entre ellas la probabilidad restante, es decir

$$p_k(u=a_j)_{j\neq c} = \frac{1-p_k(u=a_c)}{r} \qquad (12)$$

Estas curvas son sólo la estimación inicial ya que aplicando el mecanismo de aprendizaje descrito en el apartado anterior se obtendrán curvas diferentes para cada posible respuesta más ajustadas a la realidad.

El criterio de selección bayesiano descrito en el epígrafe 4 debe modificarse de manera que se calcule la esperanza matemática de la varianza para todas las posibles opciones. El criterio de selección *basado en la dificultad* requiere un valor para el parámetro de dificultad. Inicialmente este valor viene dado por el profesor, pero tras el aprendizaje automático de las curvas el valor cambia, y puede cambiar de forma diferente para cada una de las $r+1$ curvas características utilizadas. Por ello, en caso de utilizar este criterio de selección se emplea como parámetro de dificultad únicamente el asociado a la respuesta correcta.

Dado que los ítems dicotómicos no son más que un tipo especial de ítems politómicos con sólo dos opciones, en la implementación de SIETTE se utiliza internamente este tipo de ítems en ambos casos.

- ***Ítems multiopción de respuesta independiente.*** Son ítems en los que la respuesta del alumno viene dada como un conjunto de opciones entre las que el alumno debe seleccionar aquellas que considera correctas. En este primer caso se considera que las opciones son independientes entre sí es decir, que la pregunta puede descomponerse en r preguntas dicotómicas siendo r el número de respuestas posible, por ejemplo este es el caso del siguiente ítem:

***Figura 10**. Items multiopción de respuesta independiente*

Desde el punto de vista del modelo de respuesta, este caso se trata como las respuestas sucesivas a las *r* preguntas dicotómicas que lo componen. La estimación inicial de los parámetros se hace por defecto de forma global, es decir, el profesor asigna un valor a la dificultad global de la pregunta que se hace corresponder a la dificultad de cada una de las respuestas dicotómicas correctas, obteniendo las *r* curvas características de las respuestas acertadas. Las curvas restantes, para las opciones de fallo, se obtienen igualmente repartiendo la probabilidad restante en cada caso. En total se manejan $2 \times r$ curvas para un ítem de este tipo. También es posible que el profesor establezca una estimación inicial diferente para la dificultad de cada una de las opciones. Téngase en cuenta que por el tipo de control que se emplea en el interfaz, no es posible considerar como en el caso de ítems politómicos la no-respuesta como otra opción, ya que dejar en blanco una casilla es ya una respuesta.

El criterio de selección en este caso debe tener en cuenta que este tipo de ítem es equivalente a la aplicación en secuencia forzada de las *r* correspondientes preguntas dicotómicas. El criterio de selección bayesiano exige en teoría el cálculo de las 2^r combinaciones de posibles respuestas, y el cálculo de la esperanza matemática de la varianza a posteriori. Si se asume que las curvas características de cada una de las respuestas dicotómicas correctas son iguales, el cálculo se simplifica ya que solo es necesario calcular *2* opciones y ponderar en el cálculo de la esperanza matemática de la varianza según las frecuencias de cada combinación, que vienen dadas por la binomial.

Si se emplea el criterio de selección *basado en la dificultad*, es necesario precalcular la dificultad media del ítem a efectos de la adaptación. Para ello se calcula el producto de todas las curvas características de las respuestas correctas de los ítems dicotómicos obteniendo así una distribución conjunta para la que se aproxima a una curva logística mediante el procedimiento de mínimos cuadrados a fin de hallar su dificultad.

- ***Ítems multiopción dependientes entre sí.*** Un caso diferente aunque externamente similar al anterior se plantea cuando no puede suponerse independencia entre las respuestas a cada una de las opciones. Por ejemplo, supongamos el ítem:

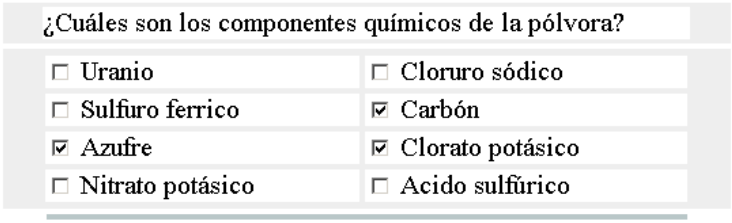

***Figura 11**. Items multiopción de respuesta conjunta*

En este caso la opción correcta es solamente una combinación de un cierto número de opciones, ocasionalmente mas de una combinación puede resultar válida. (En el ejemplo anterior la solucion es: carbón azufre y clorato potásico; aunque también es correcta la combinación carbón azufre y nitrato potásico). Este ítem se trata como un ítem politómico en el cual el número de posibles respuestas es 2^r, siendo *r* el número de opciones. La estimación inicial de las curvas la proporciona el profesor indicando las combinaciones correctas y proporcionando un valor de dificultad para cada una de ellas, iguales por defecto. El factor de adivinanza se estima inicialmente como $1/2^r$ para cada una de las combinaciones. En caso de que existan dos o más combinaciones correctas la estimación de las curvas características de cada una de ellas se obtiene, en el caso de la función logística, modificando la amplitud de la fórmula (7), es decir:

$$p_k = p_k(u=1 \mid \theta = k) = c + (\frac{1}{C} - c)\frac{1}{1 + \mathrm{e}^{-1.7\mathrm{a}\left(\left(k - \frac{K-1}{2}\right) - \mathrm{b}\right)}} \qquad (13)$$

donde C es el número de opciones correctas. La estimación inicial para las curvas características de las opciones incorrectas se obtienen distribuyendo uniformemente la probabilidad restante, de igual manera que en los ítem politómicos. En total para este tipo de ítem deben almacenarse 2^r curvas características. Al igual que en el caso anterior no se contempla la posibilidad de ausencia de respuesta.

El criterio de selección no se ve afectado, salvo en el caso en que se use el criterio de *selección basado en la dificultad*, se hayan definido múltiples alternativas correctas y se hayan asignado valores diferentes a las dificultad de cada una de ellas. En este caso el valor de dificultad que se considera es la media aritmética de los valores de dificultad de las respuestas correctas.

- ***Ítems controlados mediante programas.*** Puesto que SIETTE presenta las preguntas al usuario por medio de páginas web, una posibilidad interesante es la utilización de programas incluidos en estas paginas mediante los conocidos applets en lenguaje JAVA. Existen dos modalidades para la utilización de estos applets.

  La primera consiste en la mera incorporación de los applets a las secciones de enunciado o a las secciones de respuesta. De esta forma es posible construir preguntas en las que el enunciado se muestra dinámicamente, se pide al usuario que lo observe, y finalmente se le plantean varias opciones entre las que ha de decidir. Este mecanismo permite incluir en SIETTE temas y enunciados que midan capacidades sensoriales, visuales y/o auditivas; capacidad de atención y percepción, etc. Para el desarrollo de este tipo de preguntas, y su incorporación a un test realizado mediante SIETTE, sólo es necesario escribir de forma independiente el applet e incluirlo mediante el editor en la sección de código HTML correspondiente al enunciado o las respuestas, al igual que cualquier otro elemento multimedia.

  La segunda modalidad consiste en sustituir el mecanismo de evaluación de las preguntas que hace SIETTE incluyéndolo en el propio applet. En este caso, se plantearía al alumno un enunciado que contiene un pequeño programa que se ejecuta y muestra al usuario dinámicamente y que recoge su respuesta de forma interactiva. No son necesarias por tanto en este caso utilizar un conjunto de respuestas fijo, de entre las cuales el usuario debe elegir una; sino que el propio applet que aparece en el enunciado incluye la posibilidad de determinar si la respuesta es o no correcta. Este modo ofrece posibilidades de evaluación sensiblemente mejores que los métodos basados en selección de opciones. Por ejemplo mediante este tipo de preguntas es posible minimizar el efecto de los factores de adivinanza, controlar otros factores como el tiempo de respuesta, medir conocimientos espaciales difícilmente medibles mediante un elenco de opciones, etc.

  Este tipo de preguntas pueden integrarse en un mismo test junto con otras multirespuesta. El mecanismo de engranaje lo lleva a cabo el propio applet que se encarga de evaluar en términos de las posibles alternativas, la solución que ha obtenido el alumno mediante la interacción con el applet. Finalmente la respuesta es tratada al igual que un ítem politómico.

  La figura 12 muestra un ejemplo de este tipo de preguntas preparadas como parte de un test de conocimientos botánicos. En concreto se pretende averiguar como parte del conocimiento de esta materia, si el alumno conoce la distribución geográfica de una cierta especie. Para ello mediante un applet se muestra un mapa de Europa sobre el cual, usando una brocha verde para pintar y blanca para borrar, al estilo de los programas de dibujo, el usuario debe señalar en verde la localización de la especie. Una vez completada esta tarea, el usuario deberá pulsar el botón de "Corregir", y el applet comparará la distribución indicada por el alumno con la distribución patrón preestablecida, evaluando si la respuesta es correcta o no, de acuerdo con los márgenes de error que se consideren admisibles y que han sido programados en el applet. Una vez determinado el tipo de respuesta, el applet transmite al sistema SIETTE el resultado de la evaluación. La siguiente figura muestra este ítem justo después de pulsar el botón "Corregir".

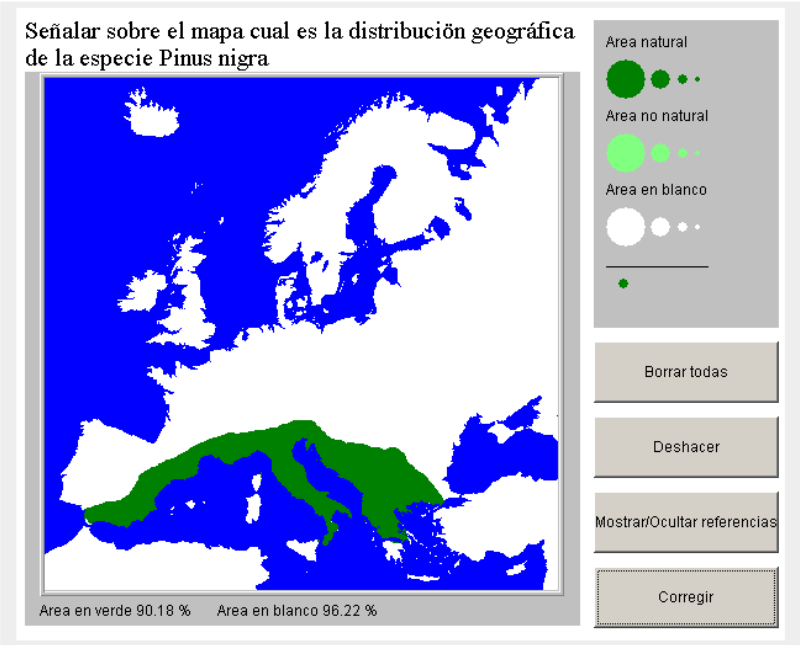

***Figura 12**. Items controlados mediante programas*

Desde el punto de vista de la implementación, la incorporación de este tipo de preguntas controladas mediante applets es muy sencilla, y resulta transparente al diseñador del test. Se ha definido una clase abstracta de la cual deben heredar los applets que realizan el control de la respuesta. Esta clase abstracta sólo obliga al constructor del applet a escribir dos funciones, una función denominada *resolver()* que es la que se encarga de mostrar al alumno la solución correcta al problema, tras su intento de resolución; y otra denominada *evaluación()* que devuelve como resultado una cadena de caracteres. Esta cadena de caracteres se corresponde a las supuestas posibles respuestas alternativas que habría dado el usuario. Al incorporar esta pregunta al test, es necesario incluir en la sección de enunciado el código correspondiente a la llamada al applet y en la sección de respuestas las etiquetas correspondientes a las respuestas simuladas. Así en el ejemplo anterior, la función de evaluación mide el porcentaje de las áreas que han sido correctamente identificadas como hábitat o no de la especie respectivamente, haya la media de estos valores y devuelve solamente una de las siguientes cuatro etiquetas "`0%-25%`"; "`25%-50%`"; "`50%-75%`"; o "`75%-100%`", que son las que el diseñador del ítem ha incluido mediante el editor como respuestas alternativas y como respuesta correcta respectivamente para este ítem politómico

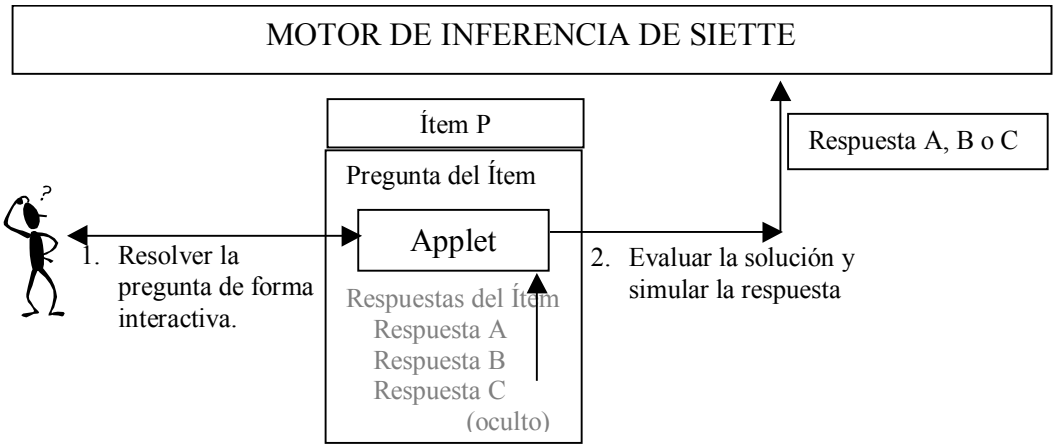

***Figura 13**. Evaluación mediante Applets*

Desde el punto de vista del modelo de respuesta y los criterios de selección los ítems controlados por applets equivalen a ítems politómicos. La estimación inicial por defecto del factor de adivinanza se considera *0*.

- ***Ítem de respuesta libre.*** El mecanismo descrito anteriormente abre un amplio abanico de posibilidades para la evaluación mediante tests. Cualquier respuesta que pueda ser evaluada mediante un programa puede formar parte de un test, transformándose en un ítem politómico. Como contrapartida, esta opción requiere ciertos conocimientos de programación. Una vía intermedia es la utilización de programas genéricos que utilicen el mecanismo descrito anteriormente, en los que el profesor mediante el editor sólo necesite configurar ciertos parámetros para especificar el enunciado y la solución. Uno de estos programas genéricos permite la inclusión de preguntas de respuesta libre.

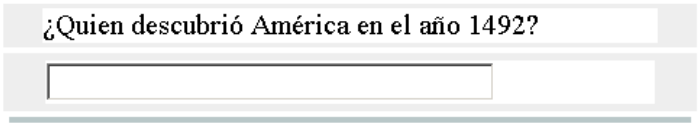

Figura 14. *Ítems de respuesta libre (texto)*

En estos casos el profesor sólo debe incluir mediante el editor, el enunciado y una expresión regular que defina las posibles respuestas. En este ejemplo una posible expresión sería `*Colón*|*Colombo*`, indicando con ello que cualquier respuesta que contenga o la secuencia de caracteres "*Colón"* o bien la secuencia "*Colombo*" será aceptada como respuesta válida, por ejemplo "*C.Colón*", o "*Cristobal Colón*" o "*Cristoforo Colombo*". Los ítems de respuesta libre se consideran ítems dicotómicos con factor de adivinanza estimado nulo.

Otra variante de este tipo de test son los de respuesta libre numérica, en los que la solución está controlada mediante una función matemática, por ejemplo para la pregunta:

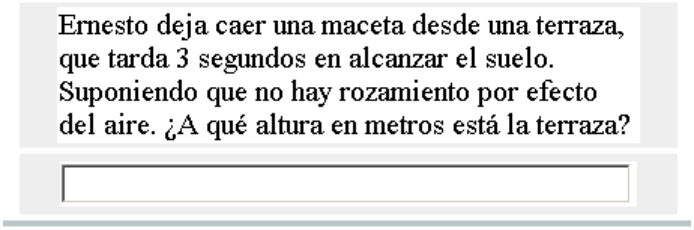

Figura 15. *Ítems de respuesta libre (números)*

el profesor debe especificar como solución: `"#(0.5*9.81*3**2)3%"` en donde indica el valor de la solución mediante una expresión que se evalúa en tiempo de ejecución y una cota del error admisible. Estas fórmulas tienen especial interés en combinación con el uso de plantillas o esquemas de preguntas que se expondrá más adelante.

Otra variante de este tipo de ítems es cuando la respuesta se divide en varias secciones de texto, en este caso el profesor debe especificar completamente el texto incluyendo las posibles expresiones válidas como respuesta. La especificación: "`<I>Completar el siguiente texto:</I><BR> "En un <<lugar>> de la <<Mancha>> <BR> de cuyo nombre no quiero acordarme..."<BR> Fragmento de El <<Quijote>>"` daría lugar al ítem:

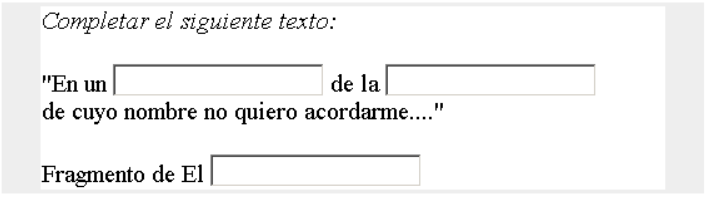

Figura 15. *Ítems de respuesta libre (formato múltiple)*

Múltiples combinaciones de expresiones regulares y fórmulas también son posibles con este tipo de ítems.

- ***Otros tipos de ítems.*** Se han desarrollado otros applets para mejorar la interfaz de usuario, aumentar la variedad y tipo de preguntas y facilitar en algunos casos la construcción de test de materias concretas, por ejemplo mediante un applet se ha implementado un ***ítem de correspondencia*** que consiste en dos columnas con *r* elementos cada una que deben enlazarse dos a dos, según una cierta relación Este ítem es realidad un ítem

politómico con *r!* posibles respuestas. La definición se realiza mediante los pares de elementos que deben enlazarse, que pueden ser texto o imágenes. El ítem se trata como un ítem dicotómico. El factor de adivinanza es en realidad *1/r!*, pero a efectos de la estimación inicial se considera nulo. Si bien el ítem acepta soluciones parciales, la evaluación sólo se considera globalmente. Hay también un ***ítem para selección de elementos de un conjunto*** Consiste en una selección interactiva posiblemente con imágenes que se desplazan mediante el ratón a una zona concreta de la pantalla. Este tipo de ítems corresponde a un ítem de selección múltiple, pudiendo ser las opciones independientes o no, pero su tratamiento en SIETTE, por el momento, es como los demás ítems basados en programas. La ventaja es que ofrece una interfaz que puede resultar útil en test informales o infantiles. Existe otro ítem que presenta una ***sopa de letras*** que consiste en localizar las palabras escondidas en una matriz de letras. Puede ser útil para test de idiomas o de terminología científica. Hay otro de ***relleno de tablas***, en el que se propone que el alumno complete una tabla con diversos números o símbolos. Esta diseñado para ser usado materias que requieran este tipo de respuestas, como por ejemplo compiladores. ***Puzzles*** en los que se propone la recomposición de una imagen previamente distorsionada, etc. Algunos ejemplos se muestran a continuación:

Figura 16. *Otros tipos de ítems*

Finalmente, otra característica importante de SIETTE es la posibilidad de utilizar plantillas o esquemas de ítems, que se instancian en tiempo de ejecución dando origen a un ítem concreto. El desarrollo de estos esquemas es independiente del tipo de ítem. Con la definición de esquemas se minimiza el riesgo de plantear preguntas repetidas, y de que el alumno pueda llegar a memorizar las preguntas existentes en el banco. El desarrollo de los esquemas se implementa actualmente mediante el uso del lenguaje PHP que es una extensión del lenguaje HTML que se interpreta en el servidor, dando lugar dinámicamente a un texto en HTML. Otros tipos de lenguajes y extensiones de HTML como PERL, JPS, etc. serían también admisibles si se encuentran instalados en el servidor. Por ejemplo, el siguiente esquema en PHP:

```
¿Cuál es el valor de x al final de este programa?  <BR><BR>
<?
   srand(date("U"));
   $randMax=getRandMax();
   $rand=Rand();
   $x =intval(doubleval($rand)*doubleval(10)/doubleval($randMax));
   echo "<CODE><PRE>";
   echo "    x=$x;<BR>";
   echo "    x++;";
   echo "</PRE></CODE>";
>
```

`<?`
`  $sol = $x+$x;`
`  echo $sol;`
`>`
	`<?`
	`  $sol = $x+1;`
	`  echo $sol;`
	`>`
		`<?`
		`  $sol = $x-1;`
		`  echo $sol;`
		`>`
			`<?`
			`  $sol = $x;`
			`  echo $sol;`
			`>`

Figura 17. *Esquema de ítem en PHP*

Daría origen a ítems en los que el valor de x es elegido aleatoriamente entre 0 y 10, por ejemplo:

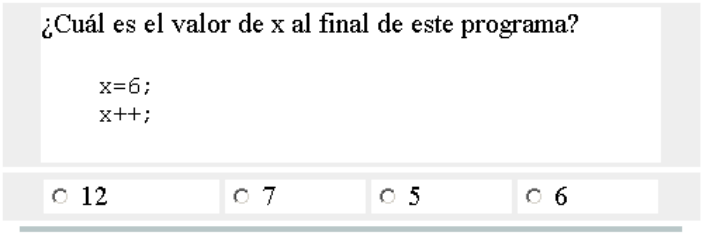

Figura 18. *Ítem generado a partir de un esquema*

Pueden emplearse esquemas para cualquier tipo de ítems y con cualquier tipo de lógica interna. Por ejemplo, en el test de Botánica que se muestra en el apartado 2, se seleccionan aleatoriamente a partir de una base de datos las imágenes de distintas especies preguntando cual es el nombre de una de ellas, o bien se selecciona aleatoriamente una especie y su correspondiente mapa y se solicita mediante el applet del ejemplo anterior que se dibuje su área de distribución.

En lo referente a las curvas características de este tipo de ítems, se asume que todas las posibles instancias de un mismo esquema tienen la misma curva característica. Esto por supuesto no es completamente cierto, y en algunos casos puede haber una mayor variabilidad que en otros. Por ejemplo en el caso del test de Botánica es mucho más difícil diferenciar especies del mismo genero que especies de distintos géneros o familias. Se deja al cuidado del profesor el garantizar que esto sea así, debiendo definir varios esquemas en caso de que considere una variabilidad de los parámetros de la curva característica. En cualquier caso nótese que el efecto de la variación en la dificultad de los ítems generados a partir de un mismo esquema es similar al supuesto del conjunto equilibrado de cuestiones definido en los apartados 5 y 6.

Los esquemas de preguntas son una herramienta potente, pero a la vez complicada. En primer lugar hacen necesario el conocimiento del lenguaje PHP o de cualquier otro que se emplee, lo cual no facilita precisamente la creación de plantillas a los no-programadores. Este problema es de difícil solución salvo para dominios concretos en los que se puedan desarrollar bibliotecas de funciones y herramientas especificas de edición para ayudar a los profesores no-programadores en la creación de esquemas. Por otra parte, es necesario un especial cuidado cuando se trabaja con esquemas, dado que ciertas instancias de un esquema pueden ser incorrectas como ítems. El esquema presentado anteriormente como ejemplo falla para x=0 y para x=1, ya que daría origen a ítems con dos alternativas iguales En esta línea la única mejora actualmente, aun sin implementar, que podría tener cierto éxito es la comprobación dinámica de que los textos de todas las alternativas sean distintos, descartando en su caso el ítem y generando otro nuevo, o seleccionando otro diferente. Sin embargo, dada la variabilidad de esquemas es imposible comprobar y solventar sistemáticamente todos los posibles casos.

8. INTEGRACIÓN EN SISTEMAS TUTORES INTELIGENTES.

La característica fundamental de los sistemas tutores inteligentes frente a los sistemas de enseñanza asistida por ordenador es que son sistemas que se adaptan al alumno ofreciendo una instrucción personalizada. (Self, 1990) Tradicionalmente estos sistemas contienen un modelo del conocimiento de la materia a enseñar, un modelo del alumno y un planificador de instrucción, que decide cual será el siguiente paso de instrucción. De poco sirve insistir sobre temas ya conocidos por el alumno, o tratar de enseñar conceptos que el alumno no está en disposición de adquirir. Para tomar estas decisiones el planificador se basa en el modelo del alumno, eligiendo el siguiente tema a enseñar y el modo de enseñarlo de manera que la instrucción sea más efectiva. Es por tanto de gran importancia para un tutor inteligente disponer de un modelo con información suficiente sobre el conocimiento que el alumno en cada momento. Este modelo frecuentemente está basado en el modelo del dominio, asignando un conocimiento a cada uno de los conceptos que componen la materia a enseñar y se conoce como *modelo de overlay*. (Polson, Ricardson, 1988). El papel de la evaluación dentro de un sistema tutor inteligente es por tanto la medición de los valores asignados a cada uno de los nodos que componen el modelo del alumno.

Hasta el momento se ha presentado SIETTE como un sistema capaz de medir un único rasgo al que se ha denominado nivel de conocimiento. Este nivel de conocimiento es un valor agregado que mide el conocimiento del alumno sobre una materia concreta. El uso de SIETTE como módulo de evaluación de un sistema tutor inteligente requiere un mayor detalle en la evaluación.

La base de conocimientos en SIETTE ha sido descrita brevemente en el apartado 2. Está compuesta por bases de conocimiento distintas para cada una de las materias a enseñar. Cada una de ellas esta formada por tres tipos de objetos: (véase fig 19)

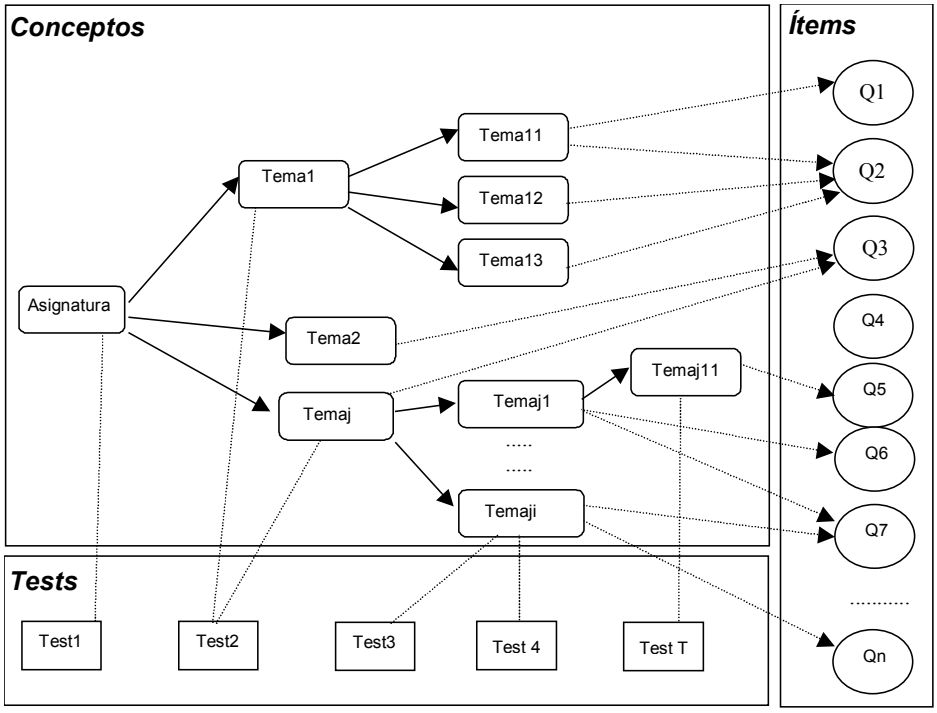

Figura 19 :*Estructura de la base de conocimientos*

- *Conceptos o temas*, que son los elementos en los que se descompone la materia a enseñar. Están estructurados jerárquicamente formando el *curriculum*. La descomposición en temas y subtemas se realiza de acuerdo a los criterios tradicionales de enseñanza, para dotar de una estructura a la materia. SIETTE puede trabajar con un número indefinido de niveles en esta jerarquía. Cada uno de los nodos finales del curriculum corresponde a un concepto único o a un conjunto de conceptos indiscernibles de cara a la evaluación. Los nodos intermedios de la jerarquía representan agregaciones de los subtemas de la jerarquía inferior según una relación de pertenencia. El modelo del alumno asocia un nivel de conocimiento para cada uno de estos temas, ya sean nodos terminales o intermedios. El curriculum lo define el profesor y su mayor o menor detalle vendrá determinado por la precisión requerida en la evaluación. Se asume la independencia entre los valores del nivel de conocimiento de cualquiera dos nodos del curriculum siempre que ninguno de ellos sea antecesor del otro.

- *Ítems.* Los ítems en SIETTE deben estar asociados explícitamente a uno o varios temas ya sean terminales o intermedios dentro del *curriculum*. Esta asociación indica que para responder correctamente al ítem es necesario el conocimiento de esos temas. La relación entre el nivel de conocimiento del tema y la respuesta al ítem viene dada por la curva característica. Hasta el momento se ha considerado que cada ítem esta asociado a un único tema, por lo que se obtienen curvas características *unidimensionales*, (véase apartado 3). En el caso en que sea necesario el conocimiento sobre varios temas para la resolución de un ítem, deben definirse curvas características *multidimensionales*, que definan la probabilidad de contestar correctamente al ítem en función de la combinación de niveles de conocimiento de los temas necesarios. En SIETTE se impone la restricción de que un ítem solo puede estar asociado a varios temas en el caso en que estos sean hermanos en la jerarquía de temas. Este aspecto se tratará mas adelante.

- *Tests*. Un test representa una sesión de evaluación. Su objetivo es obtener una estimación del nivel de conocimiento del alumno sobre uno o varios de los temas del curriculum. Por tanto, los tests se definen en función del tema o temas a evaluar. Los ítems correspondientes a un test serán los necesarios para realizar la evaluación. No existe una asociación directa entre test e ítem, salvo a través de los temas. Se impone además la restricción de que un test sólo puede estar asociado a temas que sean hermanos en la jerarquía de temas. Al asociar un test a un tema pueden prefijarse dos modos de evaluación: *agregada*, en el caso en que sólo se

requiera la evaluación de ese nodo del curriculum; o *completa*, que indica que es necesaria una evaluación exhaustiva de todos los nodos del subárbol cuya raíz es este tema.

Evaluación con ítems unidimensionales.

La primera aproximación al problema de la integración de SIETTE como módulo evaluador de un sistema tutor inteligente consiste en mantener la hipótesis de que cada ítem evalúa uno y solo uno de los conceptos o temas que componen el curriculum. En esta primera aproximación la estimación del nivel de conocimiento para cada tema puede hacerse mediante las evaluaciones sucesivas de cada uno de ellos, es decir realizar un test para cada tema. La integración en este caso es muy sencilla, ya que el sistema tutor sólo tiene que seleccionar el test asociado al tema

Desde el punto de vista procedimental, se ha implementado en SIETTE un mecanismo de integración simple que no requiere ningún compromiso especial del sistema tutor. El sistema tutor realizará una llamada a SIETTE pasándole los parámetros que definen el test, y opcionalmente el modo de visualización, si debe mostrar las respuestas o no, etc. Además pasará a SIETTE un conjunto de direcciones URL. SIETTE interpreta el número de estas direcciones como el número de niveles de conocimiento en los que se desea clasificar al alumno. Al finalizar el test, SIETTE genera una llamada a la URL correspondiente al nivel de conocimientos del alumno sobre el tema objeto del test.

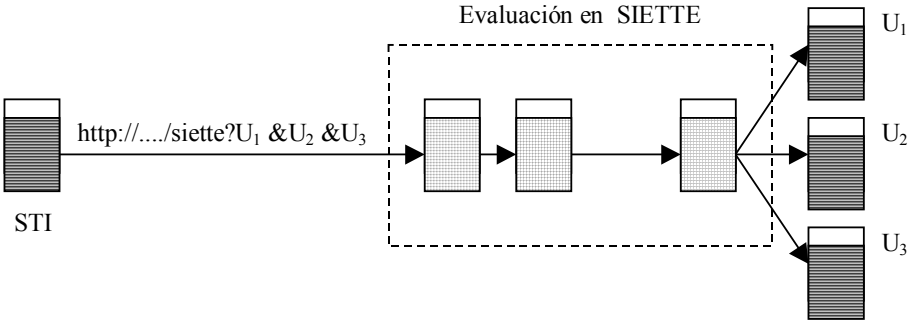

Figura 20. *Integración simple de SIETTE como componente de un sistema tutor inteligente*

Esta integración requiere un test por cada tema, lo cual no es precisamente lo mejor de cara al usuario. La alternativa es realizar un único test que mida simultáneamente el conocimiento de distintos temas. Para resolver este caso, la solución más simple consiste en suponer que cada ítem está asociado no sólo al tema para el que fue definido, sino a todos los ascendentes dentro de la jerarquía de temas. Es decir, un ítem definido para el tema T_{abc}, puede usarse también para la evaluación del tema T_{ab}, para la evaluación del tema T_a o para la evaluación de toda la materia. Esto implica disponer de $p+1$ curvas características, siendo p la profundidad del tema al que está asociado el ítem. Por el momento se ha supuesto que cada una de estas curvas es *unidimensional*. Estas curvas representan la probabilidad de contestar al ítem correctamente dado el conocimiento de cada uno de los nodos respectivamente.

El proceso de evaluación puede llevarse a cabo de forma paralela para cada uno de los nodos de la jerarquía a partir del nodo raíz. El modelo del alumno consiste en este caso en las distribuciones de probabilidad del nivel de conocimiento para cada uno de los temas del *curriculum* descendientes del nodo raíz. Formalmente, si un test está formado por los ítems $i_1,...,i_n$ siendo el vector de respuesta a estos ítems $(u_1,...,u_n)$ la estimación del nivel de conocimiento sobre el tema k, se obtiene a partir de la distribución $P(\theta_k/u_p,...,u_q)$ que es proporcional, al igual que en la fórmula (4) a

$$\mathrm{P}(\theta_k)\times\prod_{i=p}^{q}\mathrm{P_i}(\theta_k)^{u_i}(1-P_i(\theta_k))^{(1-u_i)} \qquad (14)$$

en donde $(u_p,...,u_q)$ es el subconjunto de las respuestas a los ítems asociados al tema k o a alguno de los temas descendientes de k en el *curriculum*; $P_i(\theta_k)$ representa la curva característica de respuesta al ítem i, dado el nivel de conocimiento del tema k; y $P(\theta_k)$ es la función de densidad a priori, o estimación inicial del conocimiento del alumno sobre el tema k. Nótese que con un mismo vector de respuestas es posible evaluar el nivel de conocimientos de varios nodos simultáneamente.

Por ejemplo, para evaluar el nivel de conocimiento del nodo T_a pueden emplearse todos los ítems asociados a descendientes de este nodo, como por ejemplo el ítem *i*, asociado al tema T_{abc} , utilizando las curvas características del ítem *i* asociadas al nivel de conocimiento de T_j , simultáneamente, se evalúan los niveles de conocimiento de los tema T_{ab}, y T_{abc} a partir de las curvas del ítem *i* para estos temas. Supongamos que el siguiente ítem en el test *i'* está asociado al tema T_{abd} De la misma forma, el proceso de evaluación continúa para los niveles de conocimiento asociados a los temas T_a, y T_{ab} utilizando las curvas del ítem *i'* asociadas a estos temas, y también se modifica la evaluación del nivel de conocimiento sobre el tema T_{abd} de acuerdo a la nueva evidencia.

Este modo de evaluación establece una cierta dependencia entre los valores de los niveles de conocimiento de unos temas respecto a otros. De hecho, si todos los ítems estuviesen asociados a nodos terminales en el *curriculum*, bastaría con evaluar éstos e inferir el valor del nivel de conocimiento de los nodos ascendentes mediante una función cuyos argumentos fuesen los niveles de conocimiento de sus descendientes directos. El proceso inverso no es posible, es decir, a partir de la estimación del nivel de conocimiento de un nodo, no se puede inferir el nivel de conocimiento de los nodos descendientes, salvo que se establezcan nuevas hipótesis.

Supuesto un proceso aleatorio de selección de los ítems, el número de ítem empleados en la evaluación de los nodos más altos en la jerarquía será mayor que el de los nodos terminales y por tanto la precisión en la evaluación será mayor a menor profundidad. SIETTE permite establecer un número mínimo de preguntas de cada tema para un test, con lo que se garantiza una cierta precisión.

En cuanto al mecanismo de adaptación existen varias alternativas. La más simple consiste en fijar como criterio de adaptación el mismo criterio que se usa en la evaluación del nivel de conocimiento del nodo raíz. El mecanismo de adaptación actuaría teniendo en cuenta solo la distribución del nodo raíz, y las curvas características asociadas a este rasgo. Otra alternativa más consistente, aunque computacionalmente más costosa, es calcular la influencia de la posible aplicación de un ítem en todos los vectores de conocimiento del alumno para todos los nodos de la jerarquía y establecer un criterio de mínima suma o producto de las esperanzas matemáticas de las varianzas a posteriori. Este criterio tiende a realizar preguntas de temas variados para compensar reducir todas las varianzas.

La estimación inicial de los parámetros de las *p* curvas características de un ítem, se puede realizar de distintas formas. Lo más sencillo es considerar la misma estimación inicial de todas las curvas asociadas a un mismo ítem, dejando que el mecanismo de aprendizaje se encargue de calibrarlas. Otra alternativa es que el profesor, cuando inserta el ítem en la base de conocimiento, haga una estimación directa del parámetro de dificultad de cada una de las curvas para el tema al que está asociado el ítem y todos sus antecesores en el curriculum. Esto, desde el punto de vista práctico, es un proceso bastante tedioso, ya que obliga al profesor a estimar *p+1* dificultades de la cuestión. La estimación inicial de estos valores y la convergencia del proceso de aprendizaje manteniendo la hipótesis del conjunto equilibrado es actualmente una línea abierta de investigación.

Desde el punto de vista procedimental, la integración con un sistema tutor inteligente es algo más compleja, ya que requiere del paso de parámetros desde el sistema tutor al SIETTE y desde éste al sistema tutor. Además del test a realizar, que contiene las especificaciones de los temas objeto de evaluación, los parámetros de la llamada desde el sistema tutor al SIETTE pueden ser las estimaciones a priori del nivel de conocimiento del alumno sobre cada uno de los temas. Esto no es absolutamente necesario ya que SIETTE puede utilizar distribuciones por defecto como se vio en el apartado 3, pero mejoraría la fiabilidad. Una vez realizada la evaluación SIETTE devuelve las estimaciones de los niveles de conocimiento a posteriori al sistema tutor a través de la URL que éste le proporcione en la llamada.

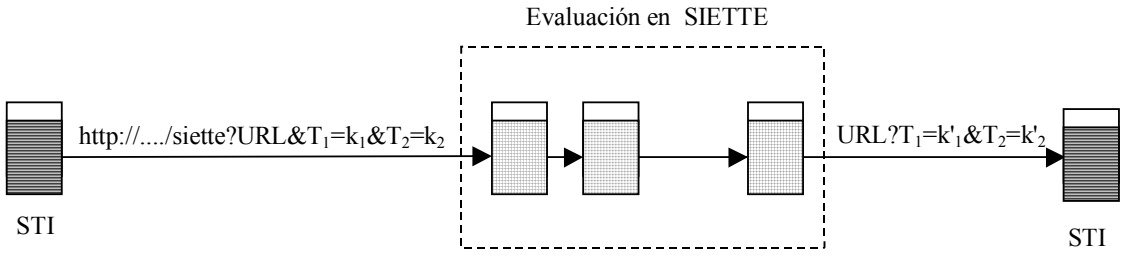

Figura 21. *Integración de SIETTE con un sistema tutor inteligente*

El principal problema de esta integración es que tanto el sistema tutor como SIETTE deben tener un curriculum común, o al menos, debe poderse establecer una correspondencia entre el modelo del alumno utilizado en el sistema

tutor y el curriculum establecido en SIETTE. Actualmente SIETTE se integra mediante esta técnica en la arquitectura MEDEA. (Trella 2001)

Evaluación con ítems multidimensionales.

Hasta ahora se ha supuesto que la respuesta a un ítem viene condicionada solamente por el nivel de conocimiento de un único tema. Esto no siempre es así, es posible encontrar cuestiones cuya resolución depende de dos o más conceptos independientes. Por ejemplo, considérese el ítem que muestra la figura 2a, donde se muestran imágenes de un pino y una pícea, y se pide al alumno que determine cual de ellos es la pícea. Si el alumno responde correctamente a la cuestión, no es posible discernir si se ha debido a su conocimiento sobre *pinos*, o ha sido por el conocimiento que tiene de las *píceas*, o quizás por el conocimiento que tiene de ambos. Si ambos temas pertenecen al curriculum, la respuesta a esta pregunta no es concluyente para la determinación del nivel de conocimientos en estos temas, pero establece una probabilidad conjunta. La presencia de estos ítems convierte a los modelos de evaluación en modelos multidimensionales.

En general, en el caso multidimensional, la curva característica de un ítem *i* viene dada por la probabilidad condicionada de responder correctamente, dada una combinación de conocimientos de los *s* temas que influyen en la respuesta, es decir $P_i(\theta_1 \dots \theta_s.) = (P(u_i=1 \mid \theta_1 \dots \theta_s.)$. En la teoría clásica de respuesta al ítem, rara vez se consideran modelos de más de dos dimensiones dada la complejidad de resolución de las ecuaciones resultantes. En este caso las curvas características de los ítems vienen dadas por funciones basadas en la distribución normal (McDonald, 1996) o logística bidimensionales (Reckase, 1996). En general, se consideran familias de curvas con un menor número de parámetros de los que cabría esperar por la agregación. Por ejemplo en el caso de la distribución normal, considerando solamente dos factores de dificultad y un único factor de adivinanza y discriminación se obtiene la familia de curvas características:

$$P(\theta,\theta') = \mathrm{c} + (1-\mathrm{c})\frac{1}{2\pi a^2}\int_{-\infty}^{\theta'}\int_{-\infty}^{\theta} e^{-\frac{(\mathrm{x}-\mathrm{b})^2-(\mathrm{y}-\mathrm{b}')^2}{2a^2}}\,dx\,dy \qquad (15)$$

Dadas las probabilidades condicionadas, para obtener los estimadores de los *s* niveles de conocimiento se aplica el procedimiento de máxima verosimilitud conjunta, o bien se calcula la densidad de la probabilidad a posteriori en el caso bayesiano $P_i(\theta_1 \dots \theta_s. \mid u_1 \dots u_n)$, que es una función *s*-dimensional proporcional a:

$$P(\theta_1 \dots \theta_s) \times \prod_{i=1}^{n} P_i(\theta_1 \dots \theta_s)^{u_i} (1 - P_i(\theta_1 \dots \theta_s))^{(1-u_i)} \qquad (16)$$

en donde $P(\theta_1 \dots \theta_s.)$ es la probabilidad conjunta a priori.

En SIETTE el cálculo de la probabilidad a posteriori es relativamente simple, dado que las curvas características se transforman en matrices *s* dimensionales de *k* componentes. En general se consideran matrices cuadradas, asumiendo que el número de niveles de conocimiento para cada rasgo es el mismo, lo que supone un total de k^s valores para cada curva. Claramente el tamaño las curvas es exponencial con el número de temas, pero desde un punto de vista práctico, el problema es abordable para valores de k^s que puedan ser procesados en un tiempo de respuesta razonable.

El criterio de selección *bayesiano* basado en la estimación de la varianza a posteriori requiere del cálculo de las distribuciones para todos los ítems antes de seleccionar el ítem a aplicar. En este caso el número de ítems es también un factor a tener en cuenta, por lo que el criterio de selección *basado en la dificultad* del ítem es claramente más eficiente.

El principal problema en este caso es el gran número de alumnos necesarios para la calibración del test usando las técnicas de estimación directas descritas en el apartado 6. Si no se establece ninguna familia de curvas, sería necesario tener evidencias del resultado obtenido para las k^s combinaciones de valores de los niveles de conocimiento, es decir para las k^s componentes de cada curva característica. Esto requiere un tamaño muy elevado de la muestra. Por otra parte, el empleo de una familia de curvas implica la aceptación de ciertas hipótesis adicionales sobre las posibles formas de las funciones de probabilidad condicionada, lo que puede resultar

excesivamente restrictivo. Una solución intermedia es aproximar la curva característica mediante hiperplanos o mediante splines *s*-dimensionales a partir de las evidencias de las que se disponga.

En cualquier caso, los modelos del alumno necesarios para los sistemas tutores inteligentes requieren muchos más componentes. La inclusión de ítems múltiples no simplifica precisamente el problema, En ciertos casos, como se ha puesto de manifiesto en el ejemplo anterior, las características del ítem obligan a realizar una definición multidimensional para mantener la corrección del modelo de respuesta, basada en la hipótesis de independencia entre los temas de un mismo nivel en la jerarquía. En otro caso se estarían falseando los resultados.

Una primera restricción que debe imponerse para controlar la explosión combinatoria que se produce al utilizar ítems multidimensionales es limitar el número de temas de los cuales pueda depender la respuesta a un ítem. Desgraciadamente, esta limitación solo afecta a las curvas características que se mantendrán menores de esa dimensión, pero no a la dimensión de la función resultante de la estimación de la probabilidad a posteriori, que vendrá dada, en el caso general, por el total de los temas que influyen en los *n* ítems del test. La condición de que los ítems multidimensionales solo pueden estar asociados a temas que tengan un mismo antecesor en del curriculum, garantiza al menos que la función podrá descomponerse en otras cuya máxima dimensión será menor que el número total de herederos

Por ejemplo, sea el ítem i_1 que depende de los temas T_1 y T_2 ; el ítem i_2 que depende de los ítems T_2 y T_3 ; y el ítem i_3 que depende de los temas T_3 y T_4. Tras la aplicación de estos tres se obtendrá una distribución conjunta de probabilidades para los niveles de conocimiento de los cuatro temas $P(\theta_1 , \theta_2 , \theta_3 , \theta_4 | u_1 , u_2 , u_3)$. Sin embargo, si se elimina de la realización del test el ítem i_2 la evaluación puede realizarse de forma independiente para cada pareja de temas, es decir se obtendrían dos distribuciones bidimensionales $P(\theta_1 , \theta_2 | u_1)$ y $P(\theta_3 , \theta_4 | u_3)$. Extrapolando esta idea, se puede limitar el número de dimensiones que se evalúan de forma conjunta mediante el análisis del grafo de interferencias entre los temas que se establece al considerar todos los posibles ítems, y aplicar un algoritmo de coloreado de grafos para eliminar del test los ítems necesarios para mantener el número máximo de dimensiones requerido. Si bien este algoritmo es a su vez exponencial, existen algoritmos eficientes que aunque no alcanzan el óptimo resultan aplicables en la práctica. Esta opción no está actualmente implementada en SIETTE siendo otra línea de investigación abierta.

El uso de ítems multidimensionales es compatible con la técnica de evaluación jerárquica en paralelo explicada anteriormente. Por otra parte, siempre es posible sustituir los ítems multidimensionales por ítems unidimensionales asociados al tema inmediato superior en la jerarquía de conceptos. Por ejemplo, la pregunta sobre pinos y píceas puede asociarse a un tema *coníferas*, que engloba a ambos y que por tanto hace innecesaria la multidimensionalidad. La diferencia es que en este caso se obtendría un modelo del alumno algo más tosco.

SIETTE es una herramienta de evaluación, y como tal los valores devueltos son datos en bruto obtenidos de las observaciones. Los sistemas tutores inteligentes disponen de herramientas para inferir de forma indirecta los valores de conocimiento de ciertos conceptos a partir de otros. Por ejemplo, es poco probable que un alumno que ha demostrado conocimiento sobre un tema no tenga conocimiento sobre otro que se considera un prerrequisito de éste. SIETTE no maneja relaciones de este tipo entre los temas, dejando estas inferencias para el sistema tutor inteligente. Otra alternativa consiste en integrar este tipo de razonamientos dentro del propio sistema de evaluación, considerando que el conocimiento de ciertos conceptos es causa de la respuesta a los ítems, y también es causa del conocimiento de otros temas o conceptos relacionados. Esto define una red bayesiana de conceptos que debe actualizarse de acuerdo a la evidencia proporcionada por la respuesta a los ítems (Millan, 2000)

9. CONCLUSIONES.

En general todo el mundo está de acuerdo en que la aplicación de la informática al campo de la educación resulta muy útil. Sin embargo, conviene plantearse cuáles son realmente las ventajas de cada sistema frente a otras opciones clásicas y de amplia raigambre. Por ejemplo, una de las principales herramientas didácticas que se utilizan en casi todas las disciplinas son textos escritos. Un texto escrito en ordenador realmente aporta pocas ventajas substanciales frente a un texto escrito en papel, su lectura resulta mucho mas incomoda y requiere del lector conocimientos adicionales de la tecnología empleada: utilizar el ordenador, el correspondiente procesador, etc. Quedan sin embargo como ventajas indiscutibles su posible mayor difusión, su bajo coste material, etc., si bien no podemos decir que se trate de ventajas substanciales, en el sentido de aportar ninguna nueva funcionalidad. Un hipertexto sin embargo, aporta una nueva forma de ver el texto, ya no como una secuencia de caracteres lineal de principio a fin, sino como

una maraña de relaciones y conceptos a través de los cuales el soporte informático permite al usuario desemvolverse con agilidad.

Pues bien, en este sentido creemos que el sistema de test por ordenador SIETTE aporta ventajas substanciales a su versión clásica correspondiente. Algunas de estas ventajas ya fueron puestas de manifiesto al hablar de los tests adaptativos informatizados frente a los tests basados en la teoría clásica de respuesta al ítem: (1) es necesario realizar un menor número de preguntas para evaluar a un alumno, y (2) a igual numero de preguntas la evaluación adaptativa es mejor, supuestas las preguntas bien calibradas. SIETTE participa de estas ventajas ya que ofrece la posibilidad de realizar tanto test adaptativos como no adaptativos. Además, la implementación discreta de la teoría resulta computacionalmente eficiente y permite abordar en ciertos casos el problema de la multidimensionalidad.

Por otra parte, al margen de estas ventajas ya conocidas, la implementación de SIETTE proporciona un valor añadido frente a los tests de papel y lápiz. En primer lugar cabe destacar el uso de plantillas o esquemas de preguntas, que facilita enormemente la construcción del banco de preguntas. Algunos dominios se prestan más que otros a este tipo de preguntas, pero en su caso ofrecen la posibilidad de creación de bancos de preguntas prácticamente ilimitados. En segundo lugar, la utilización de applets dentro de las cuestiones, la evaluación de la respuesta libre del alumno y su transformación automática en una respuesta cualitativa, permite tratar este tipo de ítems dentro del marco de la teoría formal. Algunas de estas cuestiones podrían plantearse mediante papel y lápiz, procediendo después a su corrección manual. Otras sencillamente sólo son factibles de realizar mediante el ordenador.

Hay abiertas varias líneas de investigación relacionadas con SIETTE, a las cuales se ha hecho referencia a lo largo del texto. Una línea es el estudio del modelo de respuesta basado en redes neuronales. En principio parece una idea prometedora, especialmente porque permite aprovechar directamente mecanismos de aprendizaje automático. Por el momento, los resultados obtenidos con redes de neuronas competitivas y redes de Kohonen han dado resultados comparables, aunque un poco inferiores, a los mecanismos de aprendizaje directos actualmente implementados en SIETTE y que se han expuesto en este artículo. Precisamente ésta es otra de las líneas de investigación abierta. Se han estudiado empíricamente los resultados de los mecanismos de aprendizaje y se está trabajando en la demostración formal de los mismos y la consiguiente acotación del error. También se están analizando otras hipótesis sobre el banco de ítems y una definición formal del nivel de conocimientos. En la actualidad el uso de múltiples dimensiones para la evaluación del nivel de conocimientos y su integración con el sistema MEDEA es también una línea en sus comienzos. Otra línea muy interesante sobre la que estamos trabajando es la utilización del mecanismo de preguntas y respuestas con fines tutoriales y no meramente evaluativos. Por otra parte, desde el punto de vista meramente técnico se prosigue con la implementación de una biblioteca de ítems de propósito general basados en la tecnología de applets, y de bibliotecas específicas para dominios concretos, así como en la construcción de editores específicos para estos dominios.

Actualmente se han desarrollado tests para diversas materias, relacionadas con las asignaturas que se imparten en la Escuela Técnica Superior de Informática de la Universidad de Málaga. Así por ejemplo se han creado tests para las asignaturas Procesadores de Lenguajes, Teoría de Autómatas y Lenguajes Formales, y Laboratorio de Inteligencia Artificial. También se ha desarrollado un test de Piaget para medir el desarrollo cognitivo en niños. (Arroyo,2001) Para del proyecto TREE, (Trella, 1999) se ha desarrollado un test de reconocimiento de árboles. Se está desarrollando un nuevo test de Lógica y se han hecho pruebas de tests para otros dominios. La experiencia obtenida en el uso de la herramienta ha permitido mejorarla técnicamente, aunque el sistema actual aún puede considerarse un prototipo.

El prototipo está en la dirección: http://www.lcc.uma.es/siette.

10. REFERENCIAS

Arroyo I., Conejo R., Guzmán, E., Woolf, B.P, (2001) An Adaptive Web-based Component for Cognitive Ability Estimation, en: J.D. Moore, C. Luckhardt-Redfield, W. Lewis Johnson (Eds.), *Artificial Intelligent in Education: AI-ED in the Wired and Wireless Future* (IOS Press, Amsterdam , 2001) 456-406.

Benitez R., Trella M., Conejo R., Neural networks applied to Item Response Theory, en: Neural Computation, Proceedings of the NC2000, Berlin (2000) 308-314.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's mental ability. In Lord, F. M. & Novick, M.R. (ed.) *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Brusilovsky, P. and Miller, P. (1999) Web-based testing for distance education. In: P. De Bra and J. Leggett (eds.) Proceedings of WebNet'99, World Conference of the WWW and Internet, Honolulu, HI, Oct. 24-30, 1999, AACE, pp. 149-154.

Conejo R., Millán E., Perez-de-la-Cruz J., Trella M., (2000)An empirical approach to on-line learning in SIETTE, en: Proceedings of the ITS'2000, Montreal. Springer-Verlag (2000) 57-60.

Flaugher, R. (1990). Item Pools. In Wainer H. (ed.), *Computerized adaptive testing: a primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Funahashi, K. (1998) Multilayer Neural *Networks and Bayes Decision Theory*. Neural Networks Vol11 pp 209-213.

Lord, F. M. & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (ed.), *Computer assisted instruction, testing and guidance,* pp. 139-183. New York: Harper and Row.

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

McDonald, R.P., (1996): Normal-Ogive Multidimensional Model, en van der Linden, W y Hambleton, R. (eds.) *Handbook of modern item response theory. Springer* (1996) New York.

Millán, E., Perez-de-la-Cruz, J., Suárez, E.: Adaptive Bayesian Networks for Multilevel Student Modelling, en: *Proceedings of ITS´2000*, Springer-Verlag (2000) Montreal, Canada .

Olea, J. & Ponsoda, V.: (1996) Tests adaptativos informatizados. En Muñiz, J.(ed) *Psicometría*. 1996. Madrid: Universitas.

Owen, R. J. (1975). A Bayesian sequential procedures for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association 70*, 351-356.

Polson, M.C.; Ricahrdson,J.: *Foundation of Intelligent Tutorial Systems*. Lawrence Earlbaum; Hillsdale, N.J, 1998

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

Reckase, M.D. (1996): A linear logistic multidimensional model for dichotomuos items response data, en van der Linden, W y Hambleton, R. (eds.) *Handbook of modern item response theory. Springer* (1996) New York.

Rios,A., Perez-de-la-Cruz, JL., R.Conejo: (1998) SIETTE: Intelligent Evaluation System using Test for TeleEducation , en: 4th International Conference on Intelligent Tutoring System. ITS'98. Workshops papers, San Antonio,Texas, USA (1998) .

Rios A., Millán E., Trella M., Perez-de-la-Cruz J., Conejo R.: (1999) Internet Based Evaluation System, en: Artificial Intelligence in Education AIED'99, Le Mans (1999) 387-394.

Smaminathan, H.; Gifford, J.A. (1981): *Estimation of parametres in the three parameter latent trait model.* Laboratory of psicometric and Evaluation Research. Report nº 93 Amherst, Mass. School of Education, University of Massachussets.

Santiesteban, (1990) C: Psicometría. Teoría y práctica de la construcción de test. Editorial Norma, S.A. Madrid. (1990)

Self, J. (1990) Theoretical foundation for Intelligent Tutoring Systems. AAAI/AI-ED 1990; 45.

Thissen, D. & Mislevy R.J. (1990). Testing Alghorithms. In Wainer H. (ed.), *Computerized adaptive testing: a primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D. y Steinberg, L. (1997). A Response Model for Multiple-Choice Items. Handbook of Modern Item Response Theory. 51-65.

M.Trella, D.Bueno, R.Conejo, A Web tool to help teaching morphology botany of European forestry species, en: *Advanced Research in Computer and Comunications in Education. New Human Abilities for the Networked Society. Proceedings of ICCE'99, 7th International Conference on Computers in Education*, Chiva(Japan) (1999) 1034-1040.

Trella, M.; Conejo, R. (2001). MEDEA, Una arquitectura basada en componentes para el desarrollo de sistemas tutores inteligentes en Internet. En este volumen.

Wainer, H., (ed.) (1990). *Computerized adaptive testing: a primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6,* 473-492.