SIETTE

Sistema de Evaluación de Tests para la TeleEducación

PROYECTO FIN DE CARRERA PARA LA OBTENCIÓN DEL TITULO DE INGENIERO EN INFORMÁTICA

Autora: Antonia Ríos Sanchez

Directores: José Luis Pérez de la Cruz y Ricardo Conejo Muñoz

ESCUELA TÉCNICA SUPERIOR DE INGENIEROS EN INFORMÁTICA

UNIVERSIDAD DE MÁLAGA (ESPAÑA)

Dedicado a mis padres y hermanos por su apoyo incondicional, y particularmente a mi hermana Yolanda con la que he convivido estrechamente durante la realización de este proyecto.

Mención especial merece también mi novio, informático excepcional, cuya aportación a este proyecto no sólo ha sido moral sino muchas veces práctica.

Gracias a todos por estar siempre ahí.

ÍNDICE

1. INTRODUCCIÓN

La Evaluación Adaptativa Asistida por Ordenador (CAT; Computer Adaptive Testing) es una de los avances prácticos más significativos en estrategias de evaluación. Los métodos tradicionales de evaluación mediante test, dependen del proceso estático para almacenar, manejar y analizar los datos, mientras que en los tests adaptativos este proceso es dinámico. Utilizando la información asociada a cada pregunta y a cada alumno, y aplicando el razonamiento basado en probabilidades, los sistemas CAT pueden mejorar la motivación, acortar el tiempo de test y requerir menor número de preguntas por examinando.

El razonamiento basado en probabilidades ha surgido como un acercamiento viable para la construcción, manipulación y evaluación del conocimiento en presencia de incertidumbre. Muchos de los progresos recientes en la teoría de test, se deben al estudio de las relaciones entre las respuestas dadas por un individuo a un conjunto de elementos del test y los rasgos que se suponen sobre dicho individuo, como un problema de inferencia estadística. Los principios generales de la inferencia - entre los que se encuentran los conceptos y herramientas de la probabilidad matemática - pueden ayudar a explicar las relaciones entre la evidencia y la inferencia usadas en la estimación del conocimiento del alumno y de su aprendizaje.

A su vez, los Sistemas Tutoriales Inteligentes (STI) dependen de alguna forma del modelado del alumno para poder guiar el comportamiento del tutor. Las inferencias sobre el actual conocimiento, habilidad y estrategia usada por el alumno pueden afectar a la presentación y evaluación de los problemas, la calidad del refuerzo y de la instrucción, y a la determinación del momento en el que el alumno ha completado algún conjunto de objetivos del tutorial. Sin embargo, no podemos observar directamente lo que el alumno sabe o no sabe; sino que debemos inferirlo a partir de lo que el alumno hace y no hace.

El objetivo de este proyecto ha sido desarrollar un sistema CAT de contenido equilibrado, para ser usado sobre la World Wide Web (WWW). Usando un navegador como interfaz gráfica y, simplemente, pulsado ciertos botones se podrán crear tests del tipo “verdadero-falso”, así como realizar los tests previamente creados. Además, dichos

tests no sólo destacarán por ser tests adaptativos sobre la WWW, sino que aceptarán como cuestiones ciertas plantillas que el usuario podrá definir y que facilitarán la construcción del banco de preguntas utilizado en la creación del test. A su vez, estas cuestiones y plantillas podrán incluir objetos multimedia, aprovechando de este modo la potencia que ofrece la WWW en este aspecto.

En los siguientes apartados, se explicará cómo ha sido desarrollado el sistema y cuál ha sido el punto de partida de este trabajo. Para ello, examinaremos en primer lugar los métodos tradicionales de evaluación mediante test, mostrando sus limitaciones y cómo éstas pueden ser solventadas con el uso de los tests adaptativos.

A continuación, hablaremos del motor interno de inferencia de los sistemas CAT modernos: la Teoría de Respuesta a los Elementos (IRT; Item Response Theory), y mostraremos también algunos ejemplos de su utilidad.

Seguidamente describiremos algunos de los procedimientos existentes para la selección de preguntas en los sistemas CAT, señalando las ventajas e inconvenientes de cada uno de ellos.

Aunque los sistemas IRT-CAT han sido muy útiles, existen restricciones que han impedido su extensión a una gran variedad de aplicaciones. Por ello, han surgido otros algoritmos para la construcción de tests adaptativos de los cuales, daremos una visión global.

Posteriormente, pasaremos a describir el sistema de tests implementado en SIETTE (Sistema de Evaluación de Tests para la TeleEducación) definiendo su arquitectura e incluyendo un manual de usuario en el que se indicará los pasos a seguir para la instalación y manejo del sistema.

Finalmente, mostraremos futuras directrices y aplicaciones de este proyecto, destacando su posible uso como herramienta soporte en la construcción del módulo de diagnóstico de un Sistema Tutorial Inteligente.

2. LA EVIDENCIA E INFERENCIA EN LA VALORACION EDUCACIONAL.

2.1. INTRODUCCIÓN

Este capítulo está basado en el informe técnico de Robert J. Mislevy (Mayo 1996) titulado "*Evidence And Inference In Educational Assessment*".

El problema central de la teoría de los tests puede verse como "la relación existente entre el conocimiento de un individuo sobre una tarea, y su puntuación obtenida en un test sobre dicha tarea". Se ha observado, que muchos de los progresos recientes en la teoría de test, se deben al estudio de las relaciones entre las respuestas a un conjunto de elementos del test y los rasgos que se suponen sobre el individuo, en el dominio de la materia que abarca el test, como un problema de inferencia estadística. Esta tendencia representa un progreso práctico a la hora de estar seguros para proporcionar soluciones a problemas que son formalmente intratables, tales como, cuando debe finalizar el test para cada examinando en particular y cómo clasificar las relaciones resultantes.

2.2. EVIDENCIA E INFERENCIA.

2.2.1. DATO VERSUS EVIDENCIA.

La inferencia es el proceso de razonar a partir de lo que sabemos y de lo que observamos para llevar a cabo explicaciones, conclusiones o predicciones. Además, siempre razonamos bajo la presencia de incertidumbre. La información con la que trabajamos es típicamente incompleta, inconcluyente y permite más de una interpretación. Intentaremos establecer el peso y el rango de la evidencia en lo que observamos, pero la principal pregunta que debemos contestar es *"Evidencia sobre qué"*. La distinción entre *dato* y *evidencia* puede ser vista como:

" Un dato se convierte en evidencia en algunos problemas analíticos cuando su *relevancia* sobre una o más hipótesis que están siendo consideradas, queda establecida. La evidencia es relevante en algunas hipótesis si incrementa o decrementa la verosimilitud de esas hipótesis. Sin hipótesis, la relevancia del no dato podría ser establecida."

Así, los mismos datos pueden ser concluyentes para algunas inferencias, pero apenas influir en otras; pueden abarcar un rango completo para algunas inferencias, no aportar resultados en otras; pueden constituir evidencias directas para algunas inferencias, evidencias indirectas para otras y permanecer irrelevantes, aún, para el resto.

- Las hipótesis y la comprensión de lo que constituye la evidencia sobre ellas, surgen de las variables, conceptos y relaciones entre los campos sobre los que el razonamiento tiene lugar. La valoración educacional proporciona datos tales como redacciones escritas, respuestas correctas e incorrectas en cuestionarios, presentaciones de proyectos, o explicaciones del alumno sobre su solución, al problema que se le ha planteado. Estos datos se convierten en evidencias sólo con respecto a las hipótesis sobre el alumno y su trabajo (hipótesis construidas alrededor de nociones sobre el carácter y la adquisición del conocimiento y habilidad del alumno).

No discutiremos la validez de cualquiera de esos puntos de vista. Todos ellos, son construcciones organizadas alrededor de patrones que han sido percibidos en los aspectos del aprendizaje humano. Cada uno de ellos, puede ser útil en ciertas circunstancias para mejorar el aprendizaje y la resolución, de algunos tipos de problemas. El problema que surge con esto es que, si llevamos esto a la práctica, la valoración debe realizarse teniendo en cuenta el grado de incertidumbre en toda la información que manejamos sobre un alumno.

2.2.2. CLASES DE INFERENCIA.

Algunos autores distinguen 3 tipos de razonamiento: deductivo, inductivo y abductivo. Todos ellos juegan un papel esencial y entrelazado en la valoración educacional:

➢ **Razonamiento deductivo**. El proceso de razonamiento deductivo es aquél que va de lo general a lo particular, dentro de un sistema establecido de relaciones entre las variables (de las causas a los efectos, de las enfermedades a los síntomas, del crimen a las evidencias encontradas en la escena en que se produjo, del conocimiento y habilidad del alumno a un comportamiento observable). La lógica formal incluye instancias de razonamiento deductivo concluyentes: si aceptamos que "A implica B" y aprendemos

"no B", podemos concluir "no A" con certeza. Es decir, dado un estado del sistema, ¿qué salidas puede producir éste?. En la práctica, el razonamiento deductivo es a menudo probabilístico; bajo diferentes estados, varias posibilidades son más o menos posibles, pero no están completamente determinadas.

- **Razonamiento inductivo**. Es aquél que fluye en la dirección opuesta del razonamiento deductivo y como aquél, también se produce dentro de un sistema de relaciones establecidas (de los efectos a las posibles causas, de los síntomas a las enfermedades más probables, de las soluciones o de patrones de soluciones dadas por el alumno, a la configuración del conocimiento de dicho alumno). Es decir, dadas las salidas del sistema, ¿cuáles son los estados que han podido producirlas?.

- **Razonamiento abductivo**. Es aquél que produce nuevas hipótesis, nuevas variables o nuevas relaciones entre las variables, a partir de observaciones. Es decir, es un proceso "abajo-arriba" similar al proceso de inducción, pero que se distingue de éste en que la colección de hipótesis existentes se amplía durante el proceso. Los tests evidenciales de estas nuevas hipótesis, son inferidos de manera deductiva, a partir de la nueva hipótesis.

Las teorías y explicaciones de un campo sugieren la estructura a través de la cual el razonamiento deductivo fluye. El razonamiento inductivo y abductivo dependen igualmente de las mismas estructuras, aunque el recorrido de las mismas en busca de los objetivos es diferente.

2.2.3. RAZONAMIENTO BASADO EN PROBABILIDAD.

Nosotros no construimos modelos basados en probabilidades en la mayoría de razonamientos que hacemos. Continuamente razonamos deductivamente, inductivamente y abductivamente, para estar seguros, pero no a través de modelos formales explícitos. No lo hacemos, principalmente porque usamos heurísticos que son suficientes para nuestro propósito diario. Otra razón importante, es que la mayoría del razonamiento, lo hacemos sobre dominios de los cuales conocemos información.

Heurísticos, hábitos, reglas, pruebas y los típicos procedimientos operativos, guían en la práctica el conocimiento sobre un dominio; más o menos en respuesta a lo que parece haber sucedido en el pasado, y que parece estar relacionado con el problema a resolver. Esta maquinaria inferencial evoluciona y se relaciona con los problemas,

conceptos, restricciones y la metodología de un campo. Las dificultades llegan cuando los problemas inferidos son tan complejos que los heurísticos normales fallan o cuando aparecen problemas nuevos. Es, en estas situaciones, cuando los sistemas de inferencia desarrollados formalmente, proporcionan gran valor.

Dados los conceptos y relaciones claves, objetivos inferidos y datos, ¿cómo se debería llevar a cabo el razonamiento?. Una vertiente tradicional en la teoría de tests, ha surgido con el tiempo: **la probabilidad matemática**. Para nuestros propósitos, los elementos esenciales son un conjunto especificado de salidas (también llamado muestra del espacio del problema); un conjunto de variables (que determinan cómo de probables son las salidas), y una función (que especifica las probabilidades de los eventos o subconjuntos del espacio muestral), dando valores a los parámetros.

Así, dado los valores de los parámetros, podemos expresar la probabilidad de un evento, y compararlo con cualquier otro; y dado un evento, podemos expresar la verosimilitud del valor de un parámetro al compararlo con el resto de valores de los parámetros.

Como discutiremos posteriormente, la probabilidad matemática proporciona herramientas para combinar la evidencia dentro de una estructura.

3. SISTEMAS TUTORIALES INTELIGENTES

3.1. INTRODUCCIÓN

La inteligencia artificial en la educación asistida con ordenador surge con los sistemas tutoriales inteligentes, que no son sino una evolución de la enseñanza asistida por ordenador. Tres aspectos caracterizan a estos sistemas inteligentes. Primero, el dominio o problema a tratar debe ser conocido por el sistema informático, lo suficientemente bien como para poder generar inferencias o resolver problemas en ese dominio. Segundo, el sistema debe ser capaz de enseñar ese conocimiento al alumno. Tercero, la estrategia tutorial o de enseñanza debe tener la inteligencia suficiente para implementar estrategias que reduzcan la diferencia de rendimiento entre el experto y el estudiante.

Como partes de un Sistema Tutorial Inteligente (STI) tenemos:

1. El **módulo experto** que contiene el conocimiento del dominio.
2. El **módulo de diagnóstico del alumno** que analiza lo que el estudiante sabe.
3. El **módulo instructor** que identifica las deficiencias en el conocimiento, para seleccionar estrategias que enseñen ese conocimiento del modo adecuado.
4. El **entorno de instrucción** y la **interfaz humano-maquina** que canalizan la comunicación del tutorial.
5. Además de esos componentes, los temas de **implementación** y **evaluación** también son importantes, al igual que encontrar respuesta a: ¿Cuándo, dónde y cómo deberían ser usados estos STI? ¿Qué efectividad tiene el STI y cómo se analiza su calidad?.

3.2. EL MÓDULO EXPERTO

El módulo experto es la parte de un tutor que provee el dominio del conocimiento. La mayor lección que la comunidad de la inteligencia artificial ha aprendido de todas las investigaciones en sistemas expertos es que cualquier módulo experto debe tener una abundancia de conocimiento específico y detallado, derivado de la gente que tiene años de experiencia en un dominio particular. Consecuentemente, es

necesario un gran esfuerzo para descubrir y codificar el conocimiento del dominio. La gran cantidad de conocimiento en dominios complejos junto con la relación de ese conocimiento, hace que el diseño y desarrollo del módulo experto deba ser la tarea más complicada al desarrollar un STI. Los diseñadores de tutores inteligentes pueden encontrar difícil descubrir y codificar todo el conocimiento necesario del dominio. Por lo tanto, investigar como codificar el conocimiento, y como representar tal experiencia en un STI, continua siendo el objetivo principal en el desarrollo del módulo experto.

3.3. EL MÓDULO DE DIAGNÓSTICO DEL ALUMNO

Muchos STIs infieren un modelo del actual conocimiento del alumno sobre el problema y usan ese conocimiento para adaptar la instrucción o las necesidades particulares del alumno. La estructura del conocimiento que proporciona el estado actual del alumno es el *modelo del alumno,* y el proceso de razonar para desarrollarlo se llama *diagnóstico del alumno*. Las salidas del módulo de diagnóstico del alumno pueden usarse para diferentes propósitos, tales como avanzar a través de currículums seleccionados, enseñar u ofrecer consejos no solicitados, generar nuevos problemas y adaptar conjuntos de explicaciones. Los expertos discuten la necesidad de:

a) mejorar la cantidad de conocimiento disponible sobre el alumno (Ancho de Banda),

b) identificar los tipos de conocimientos que se van a enseñar, y

c) calcular las diferencias entre el experto y el alumno.

Diseñar el módulo de diagnóstico del alumno es un gran riesgo, y consecuentemente presenta un amplio rango de posibilidades a investigar, como las siguientes: ¿Qué nivel de detalle debe tener la descripción del conocimiento del alumno? ¿Qué modelos de aprendizaje pueden diseñarse como una superestructura para los algoritmos de diagnóstico? ¿Cómo debería la comunidad de investigación en inteligencia artificial dirigir la tecnología de los sistemas expertos hacia la tecnología de los STI?.

3.4. EL MÓDULO DE INSTRUCCIÓN Y CURRÍCULUM

Un STI debería tener tres características tutoriales: a) controles sobre la representación del conocimiento de la instrucción para seleccionar temas y guiar su presentación dependiendo del alumno; b) capacidad para responder a las cuestiones de

los alumnos sobre los objetivos y contenidos de la instrucción y c) estrategias para determinar cuándo necesitan ayuda los alumnos para ofrecerles la información necesaria. Separar la instrucción del contenido del experto es el desafío que existe para el diseño del módulo de instrucción aunque evidentemente el conocimiento y el proceso de aprender están relacionados con la forma de enseñar. Quizás sea menos obvia, la interacción entre los contenidos específicos y la estrategia de instrucción.

3.5. EL ENTORNO DE INSTRUCCIÓN

Un entorno de instrucción consiste en esos elementos de un STI que soportan lo que el alumno esta haciendo: situaciones, actividades y herramientas que provee el sistema para facilitar el aprendizaje.

Las actividades y herramientas presentadas al alumno en un STI siempre reflejan una filosofía educacional subyacente. Como los ordenadores cada vez son más rápidos y los investigadores de STI y educadores son cada vez más creativos, es fácil crear una experiencia educacional más abierta, robusta y efectiva. Además debería ser diseñado para que los estudiantes se sintieran auto-monitorizados, permitiendo a los alumnos asumir responsabilidades de su propio aprendizaje.

Se puede asegurar que el éxito en la construcción de entornos de instrucción dependerá en gran medida de lo bien que esté diseñada la interfaz humano-máquina.

3.6. HACIA SISTEMAS EDUCACIONALES BASADOS EN EL CONOCIMIENTO

Aún no se han realizado los suficientes STI como para tener una tecnología comprensible, se necesita más experiencia y más STI para la exploración de todas las posibilidades. Sin embargo, la educación y las comunidades de entrenamiento sólo podrán esperar buenos resultados cuando aparezca una tecnología formal para los STI. Así que no sólo se necesitan nuevos STI para exploración, sino para determinar de una forma generalizada como desarrollar los STI. Este desarrollo no será una tarea simple. Lo que sí está claro es que tales sistemas necesitarán siete tipos de habilidades como mínimo. Estas habilidades pertenecen a los componentes que deben ser integrados como el fundamento de un STI:

1) Habilidades de contenido en el módulo experto.

2) Habilidades de diagnóstico (determinan lo que el alumno sabe y lo que debe aprender) en el módulo de diagnóstico del alumno.
3) Destreza instructiva y curricular en el módulo instructor.
4) Destreza para crear los entornos de instrucción
5) Habilidad para crear la interfaz humano-máquina.
6) Habilidad de implementación y
7) Destreza de evaluación.

Estos componentes comprenden la anatomía de los STI, y juntas ofrecen un modelo conceptual básico para diseño, desarrollo, distribución y evaluación de máquinas tutoras. No es fácil integrar todo ese conocimiento en un único sistema. Los STI prometen no ser sólo ayuda para personas que quieren aprender a realizar tareas complejas mejor, sino también ayudaran a revelar cómo aprende la gente.

4. TEORIA DEL TEST.

4.1. INTRODUCCIÓN.

El propósito de este apartado es examinar los recientes desarrollos en el método de evaluación mediante test con el uso de las computadoras. También examinaremos el uso dominante de los tests adaptativos como medida educacional, en un intento de pronosticar algunas directrices futuras en la investigación de los métodos de evaluación mediante test.

El objetivo de un test adaptativo asistido por ordenador (CAT: Computerized Adaptive Tests) es usar el menor número de preguntas para determinar el nivel de conocimiento del alumno en un determinado dominio. Para ello, los CAT evitan plantear preguntas que son demasiado fáciles o demasiado difíciles para un determinado alumno, y en su lugar, eligen preguntas cuya dificultad es más cercana, al nivel de conocimiento del alumno. Por tanto, en primer lugar, las preguntas del banco de los CAT deberían medir todas la misma cosa. En segundo lugar, se supone que los elementos del banco son independientes unos de otros, es decir, la probabilidad de que el alumno responda correctamente a una pregunta no debería afectar al orden en que las preguntas son preguntadas. Finalmente, las preguntas del banco son seleccionadas sin reemplazamiento, es decir, una vez que una pregunta es presentada al alumno, ya no se vuelve a usar durante esa sesión de test.

Muchas de las aplicaciones más recientes de los CAT hacen uso de la teoría IRT, para la selección de las preguntas del test a generar y para estimar el nivel de conocimiento alcanzado por el alumno y su precisión. Estas estimaciones pueden ser usadas en conjunción con ciertas estrategias de evaluación para facilitar ciertas decisiones educacionales.

Tras el modelo IRT surgieron otras alternativas como SPRT y EXSPRT, debido a la complejidad matemática y conceptual de dicho modelo, que limitaba el uso de los tests adaptativos en entornos educacionales.

Tanto el modelo IRT como otras alternativas existentes en la teoría de tests adaptativos serán descritas en sucesivas secciones, centrándonos a continuación en las

limitaciones de los métodos tradicionales de evaluación mediante test, y en la descripción global de los componentes claves de todo CAT.

4.2. **TESTS ADAPTATIVOS ASISTIDOS POR ORDENADOR.**

4.2.1. MODELOS CLÁSICOS DE TEST. LIMITACIONES.

Los métodos tradicionales de evaluación mediante test, dependen del proceso estático para almacenar, manejar y analizar los datos, mientras que en los tests adaptativos este proceso es dinámico.

Los modelos clásicos de test se han centrado en la creación de diferentes test para diferentes conjuntos de estudiantes. Cada test es diseñado de modo que un estudiante con un conocimiento medio en la muestra, debe obtener una puntuación media. Tales modelos tienen 3 limitaciones, en el valor de los datos que se obtienen del test.

La primera limitación de los modelos tradicionales (test sobre papel) está relacionada con las estadísticas extraídas de las respuestas de los estudiantes. De este modo, las estadísticas obtenidas a partir de una muestra de estudiantes, son solamente aplicables a aquellos estudiantes que tienen características similares a los de la muestra.

La siguiente limitación es referente a la comparación de los estudiantes de una muestra con aquellos estudiantes pertenecientes a otra muestra. Esto es así, ya que resulta muy difícil comparar las estadísticas de distintos estudiantes, a menos que éstos, sean evaluados por tests equivalentes. Tales comparaciones requerirán procedimientos complejos ya que no existe una relación lineal entre las estadísticas de los resultados en test no equivalentes. Por otro lado, es extremadamente difícil, y raramente ocurre, que el diseñador del test desarrolle tests equivalentes.

La última limitación del modelo clásico está relacionada con la efectividad del test, es decir, si el test mide de manera precisa el verdadero conocimiento del estudiante. Aunque es sabido que la mayoría de los tests cometen un error entre lo que el alumno realmente sabe y lo que los resultados del test reflejan, en el modelo clásico se supone que este error es el mismo para todos los alumnos. Así, aunque el nivel de conocimiento determinado en el test es erróneo, se supone que la valoración del estudiante en comparación con otros estudiantes de la muestra es exacta, cosa que no ocurre.

Concluyendo, en el modelo clásico o tradicional, cada test tiene un rango en el cual puede, de manera más precisa, diferenciar los distintos niveles de conocimiento de los estudiantes. Estudiantes con niveles de comprensión dentro del rango efectivo del test, serán valorados con mayor exactitud. Aquellos estudiantes con un nivel de conocimiento por encima o por debajo de este rango, son más propensos a ser valorados con un nivel menos exacto al real.

Investigaciones sobre estas limitaciones sugieren que éstas pueden ser solventadas con el modelo **IRT** (Teoría de Respuesta a los Elementos: Item Response Theory).

4.2.2. COMPONENTES DE UN PROCEDIMIENTO DE EVALUACIÓN ADAPTATIVA MEDIANTE TESTS.

Un test adaptativo asistido por ordenador (CAT) posee los siguientes componentes: (a) un modelo de respuesta asociado a cada pregunta, (b) un banco de preguntas posibles, (c) un nivel de entrada, (d) una regla de selección de las preguntas, (e) un método de puntuación y (f) un criterio de finalización.

- **Modelo de respuesta asociado a cada pregunta.**

Dentro del contexto de la teoría IRT (que describiremos en la siguiente sección), los CAT pueden ser actualmente implementados usando 1, 2 0 3 parámetros. La elección del modelo a usar debería depender de la naturaleza de las preguntas que constituirán el test (multi-elección, verdadero-falso, palabras clave, etc.) y de la forma de las posibles respuestas a las preguntas. El resto de componentes mostrados a continuación, son independientes del modelo elegido.

- **Banco de preguntas posibles.**

Constituye uno de los elementos esenciales para la creación de un CAT. Las preguntas deben estar calibradas de alguna forma (en el caso de la teoría IRT, dicha calibración consistirá en estimar los parámetros asociados a cada pregunta), y usando una métrica común con los procedimientos usados. Se consideran satisfactorios bancos con más de 100 preguntas. Es esencial para todas las aplicaciones de los CAT que los niveles de dificultad de las preguntas y respuestas (elementos del test) abarquen el rango completo de los niveles existentes en el dominio del que se va a evaluar al alumno. Las

medidas más eficientes para diagnosticar el nivel de conocimiento de un alumno son aquellas que usan preguntas con altos índices de discriminación.

Investigaciones recientes, plantean cuestiones sobre si los procedimientos adaptativos crearán contenidos prefijados en los tests generados, ya que diferentes preguntas son administradas para diferentes alumnos. En un dominio heterogéneo, esta pregunta debería dirigirse hacia la estructuración del banco y asegurar que diferentes áreas de contenido del dominio, están igualmente representadas en los diferentes niveles de dificultad del banco. Además, si es posible, las diversas áreas de contenido deberían también estar balanceadas con respecto a sus índices de discriminación.

- **Nivel de entrada.**

En los tests adaptativos es posible empezar con preguntas de diferentes niveles de dificultad, para diferentes estudiantes. Ya que el nivel de dificultad de las preguntas seleccionadas se moverá alrededor del nivel de conocimiento del alumno, a medida que el test progresa. Un error en el nivel de entrada no afectará seriamente al resultado puesto que este se irá modelando conforme el test avanza. En cambio, la exactitud del nivel de entrada reducirá el número de preguntas que se necesitan para alcanzar con precisión dicho resultado.

- **Regla de selección de las preguntas.**

Debido a la filosofía de los CAT, hay que hacer la pregunta adecuada a cada alumno, dependiendo de su nivel de conocimiento. Para alcanzar este fin, debemos fijar una regla de selección de preguntas, que sea adecuada con el objetivo del test a realizar. Existen muchos y variados métodos entre los que destacan dos procedimientos: Máxima información y Bayesiano (Owen, 1975). Ambos procedimientos conllevan una búsqueda completa en el banco de preguntas no presentadas aún al alumno, para seleccionar una de ellas. En la selección de la pregunta que aporta la máxima información, se selecciona aquella que proporciona la máxima información sobre la última estimación del nivel de conocimiento del alumno (θ). En la selección Bayesiana de la pregunta, se elegirá aquella que minimiza la varianza a posteriori esperada de la estimación, θ. Cómo cantidad de información y varianza a posteriori son funciones relacionadas, en muchos casos, con ambos métodos se seleccionará un subconjunto similar de preguntas para un estudiante determinado.

Independientemente del algoritmo elegido, éstos deben completarse con algoritmos de restricciones, para resolver así, problemas como el de los contenidos no equilibrados. En este caso, se ha de modificar la regla de selección para asegurar que todas las áreas de contenido sean cubiertas. Bajo estas circunstancias, la regla de selección se ha de aplicar separadamente a cada una de las áreas definidas.

Estos métodos junto con otros menos óptimos, serán presentados en sucesivas secciones de este capítulo.

- **Método de Puntuación.**

Los niveles de conocimiento pueden ser estimados por distintos métodos, destacando el método de Máxima Verosimilitud o bien, el método de Actualización Bayesiana.

El método de actualización Bayesiana proporciona una estimación única (θ) del conocimiento del alumno para las respuestas a simples preguntas o para patrones de respuestas que son todas correctas o todas incorrectas. El método de la Máxima Verosimilitud, no puede proporcionar estimaciones únicas bajo estas circunstancias. Por otro lado, las estimaciones Bayesianas tienden hacia la estimación θ a priori. Esta regresión es más fuerte en los tests relativamente cortos.

Las estimaciones de Máxima Verosimilitud tienden a ser totalmente imparciales. Un compromiso práctico sería usar la puntuación Bayesiana durante la primera parte del test y a partir de ahí, realizar estimaciones con la Máxima Verosimilitud.

Ambos métodos de puntuación proporcionan estimaciones del error de la varianza, las cuales pueden usarse para crear intervalos de confianza o credibilidad respecto a la estimación θ realizada. Estos intervalos sirven para delimitar la precisión de la estimación del nivel alcanzado: el menor intervalo produce la estimación más precisa. Dichos intervalos de confianza son muy útiles en los tests adaptativos.

- **Criterio de terminación.**

Una importante característica de los CAT es que éstos sólo continúan hasta que éstos pueden decidir el nivel de conocimiento actual del alumno. Pueden adoptarse varios criterios de finalización, según el propósito del test; aunque bien es cierto que dichos criterios estarán íntimamente relacionados con los algoritmos de selección de preguntas y con los métodos de puntuación elegidos.

4.3. IRT Y EVALUACIÓN ADAPTATIVA ASISTIDA POR ORDENADOR

4.3.1. INTRODUCCIÓN

La Evaluación Adaptativa Asistida por Ordenador (CAT; Computer Adaptive Testing) es una de los avances prácticos más significativos. Utilizando la información de sus patrones de respuesta para seleccionar adaptativamente las preguntas a plantear a los examinados, los sistemas CAT pueden mejorar la motivación, acortar el tiempo de test y requerir menor número de preguntas por examinando, todo esto sin sacrificar la fiabilidad de la medida. El motor interno de inferencia de los sistemas CAT modernos es la Teoría de Respuesta a los Elementos (IRT; Item Response Theory).

Aunque los sistemas IRT-CAT han sido muy útiles, dos restricciones han impedido su extensión a una variedad más grande de aplicaciones. Estas restricciones son:

- El ámbito limitado de tareas que pueden ser utilizadas sin violar, de forma seria, las premisas sobre la independencia condicional asumidas por el método IRT.
- Las capacidades limitadas de IRT para tratar simultáneamente los aspectos múltiples e interdependientes del conocimiento o nivel del alumno.

Los modelos gráficos (GM; Graphical Models), llamados también Redes de Inferencia Bayesiana (BIN; Bayesian Inference Networks) proporcionan un lenguaje para describir dependencias multivariables complejas. La perspectiva del modelo gráfico extiende el marco de inferencia de IRT-CAT para acomodarlo a tareas más ricas y modelos de estudiante más complejos.

4.3.2. IRT (ITEM RESPONSE THEORY) Y CAT (COMPUTERIZED ADAPTIVE TESTING)

Un modelo IRT expresa la tendencia de un examinando a responder correctamente, o recibir puntuaciones altas en una colección de preguntas de test, en términos de la variable de conocimiento no observable θ. Se parte de la hipótesis de que las respuestas son independientes, condicionadas por θ y otros parámetros que expresan las características de las preguntas, como puede ser su dificultad o su índice de

discriminación. Un ejemplo simple es el modelo de Rasch para n preguntas dicotómicas:

$$P(x_1,\ldots,x_n \mid \theta, \beta_1,\ldots,\beta_n) = \prod_{j=1}^{n} P(x_j \mid \theta, \beta_j) \qquad (1)$$

donde x_j es la respuesta a la pregunta j (1 para correcta, 0 para incorrecta), β_j es el "parámetro de dificultad" de la pregunta j y

$$P(x_j \mid \theta, \beta_j) = \exp\left[x_j(\theta - \beta_j)\right] / \left[1 + \exp(\theta - \beta_j)\right] \qquad (2)$$

A fin de seleccionar preguntas y puntuar a los examinandos en aplicaciones típicas, se obtienen estimaciones puntuales de los parámetros de la pregunta $(\beta_1,\ldots,\beta_n)$, abreviadamente **B**, a partir de grandes muestras de respuestas de los examinandos y se tratan como conocidos. Más adelante se discutirá el modelado de fuentes alternativas de información y de la incertidumbre restante sobre **B**.

Una vez observado un vector de respuesta $x = (x_1,\ldots,x_n)$, la ecuación (1) es interpretada como una función de verosimilitud para θ, digamos $L(\theta \mid x, B)$. El MLE (Estimador de Máxima Verosimilitud; Maximum Likelihood Estimator) $\hat{\theta}$ maximiza $L(\theta \mid x, B)$; su varianza asintótica puede ser aproximada por la función recíproca de la función de información de Fisher, o el valor esperado de la segunda derivada de $-L(\theta \mid x, B)$, evaluada en $\hat{\theta}$. La inferencia Bayesiana está basada en la distribución a posteriori $p(\theta \mid x, B) \propto L(\theta \mid x, B)p(\theta)$, la cual puede ser resumida en términos de: la media a posteriori $\bar{\theta}$ y la varianza a posteriori $Var(\theta \mid x, B)$.

Los exámenes de forma fija (por ejemplo, todos con el mismo número prefijado de preguntas) obtienen diferente exactitud en la evaluación para valores diferentes de θ, con mayor precisión cuando θ está próxima a las dificultades de las preguntas. El método CAT proporciona la oportunidad de ajustar el nivel de dificultad para cada examinando. El test se desarrolla de forma secuencial, seleccionando cada pregunta sucesiva $k+1$ de forma que sea informativa sobre el parámetro θ del examinando en función de las respuestas dadas a las primeras k preguntas, ó $x^{(k)}$. Una aproximación muy común es evaluar $\bar{\theta}$ después de cada respuesta, y posteriormente seleccionar la siguiente pregunta de la base de preguntas que proporcione un valor grande de la

función de información de Fisher en las cercanías de $\bar{\theta}$. Una aproximación Bayesiana selecciona la siguiente pregunta como aquella que minimiza la varianza a posteriori esperada, o $E_{xj}\left[Var\left(\theta \mid x^{(k)}, x_j, B^{(k)}, \beta_j\right) \mid x^{(k)}, B^{(k)}\right]$ (Owen, 1975). Restricciones adicionales para la selección de preguntas, tales como contenido de la pregunta y formato, son abordadas más adelante. El test termina cuando se ha obtenido la precisión de medida deseada o se ha presentado un número de preguntas predeterminado.

4.3.2.1. Ejemplo: Un modelo IRT.

El modelo Rasch para tests de preguntas dicotómicas, se usa para inferir del nivel de conocimiento general del alumno, sobre el dominio particular al que pertenece el test. Dicho modelo postula que las respuestas a n preguntas del test de un dominio, son independientes dados los parámetros que caracterizan el conocimiento general del alumno para dar respuestas correctas (denominado θ) y la dificultad de cada pregunta del test (β_j denota el parámetro dificultad de la pregunta j). La figura 1 muestra las probabilidades de responder correctamente a 3 preguntas, con parámetros de dificultad –1, 0, y +1, como una función de θ. Valores bajos de θ indican bajas oportunidades de responder correctamente a la pregunta, mientras que valores altos de θ indican oportunidades altas; como porcentajes determinados por los parámetros de la pregunta.

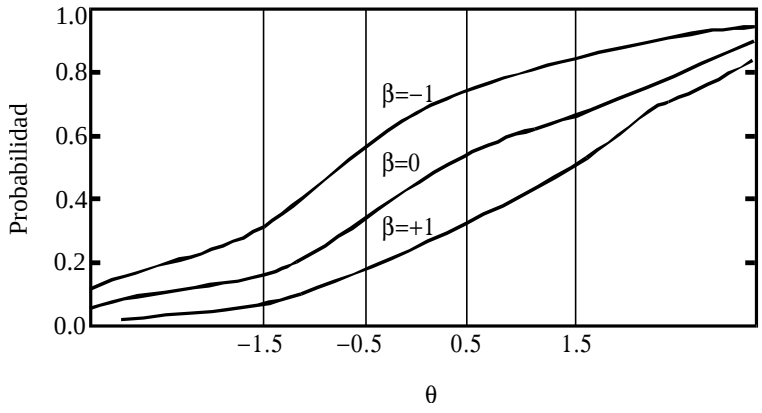


Figura 1. Probabilidad de la respuesta correcta, condicionada sobre θ, para preguntas con β = -1, β = 0, y β = 1.

La figura 2 muestra las relaciones expresadas en la ecuación (1), entre las variables pertenecientes a un alumno, como un grafo acíclico dirigido (GAD). Cada nodo representa una variable (el nivel de conocimiento del alumno) y 3 preguntas X_1, X_2 y X_3. Una flecha entre nodos representa una relación de probabilidad condicional

entre las variables, la dirección de la flecha significa sobre qué variable se está condicionando (de "padres" a "hijos", en la terminología de los GAD). La ausencia de flechas entre las variables ($X's$) representa la independencia condicional; estas variables sólo estarán relacionadas a través de θ. Para cualquier X_i la distribución de probabilidades es modelada dependiente de θ y de β_j.

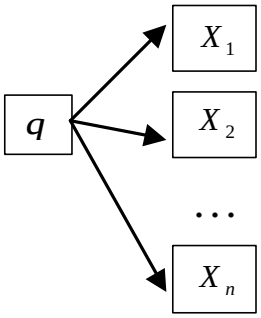


Figura 2. Gráfico acíclico dirigido para el ejemplo IRT.

Las ecuaciones (1) y (2) representan la inferencia deductiva a partir de θ y de las $\beta's$ para medir la creencia sobre las $X's$. Alternativamente, si se dan valores particulares de θ y $\beta's$, las ecuaciones (1) y (2) podrían usarse para asignar probabilidades a las hipótesis sobre las respuestas que se podrían observar, tales como: "La respuesta a la pregunta 1 será 0 en lugar de 1 con una probabilidad de 0.8", etc.

Las flechas en la figura 2 indican la estructura de las relaciones, pero no su intensidad. Supondremos por simplicidad, que θ puede tomar sólo 4 valores: -1.5, -0.5, 0.5 y 1.5, y que los valores de β para tres preguntas son: -1, 0 y 1 respectivamente. La tabla 1 nos da las probabilidades de responder correctamente a cada pregunta dado el nivel de conocimiento θ, previamente calculado a partir del teorema de Bayes, que dice:

$$p(z \mid x) = \frac{p(x \mid z)\, p(z)}{p(x)} \qquad (3)$$

donde $p(x)$ es el valor esperado de x sobre todos los posibles valores de z, o expresado matemáticamente:

$$p(x) = E\left[p(x \mid z)\right] = \begin{cases} \int p(x \mid z)p(z)d(z) & z \quad continua \\ \sum p(x \mid z)p(z) & z \quad discreta \end{cases} \qquad (4)$$

con la integral o suma realizada en el rango de valores admisible para z.

Podemos observar en la ecuación (3) que los términos que cambian la creencia sobre una hipótesis, desde $p(z)$ a $p(z \mid x)$, son denominados verosimilitudes, $p(x \mid z)$; es decir, las probabilidades relativas del dato observado, dado cada uno de los posibles estados que podrían haberse producido. Mientras que las expresiones $p(x \mid z)$ generan un razonamiento *deductivo* sobre las posibles respuestas a una pregunta x, suponiendo conocido el valor de z, la misma expresión genera también un razonamiento *inductivo* sobre la verosimilitud o credibilidad del posible valor de z, una vez que se ha observado el valor de x.

Desde una perspectiva de estadística Bayesiana, las verosimilitudes caracterizan completamente el peso y la dirección del valor evidencial que las observaciones sostienen sobre las hipótesis.

Este último punto resalta la caracterización de la creencia y el peso de la evidencia bajo el paradigma de la probabilidad matemática:

- Antes de observar un dato, la creencia sobre él, según el conjunto de los posibles valores que puede tomar, se expresa como una distribución de probabilidad (densidad), llamada distribución a priori $p(z)$.

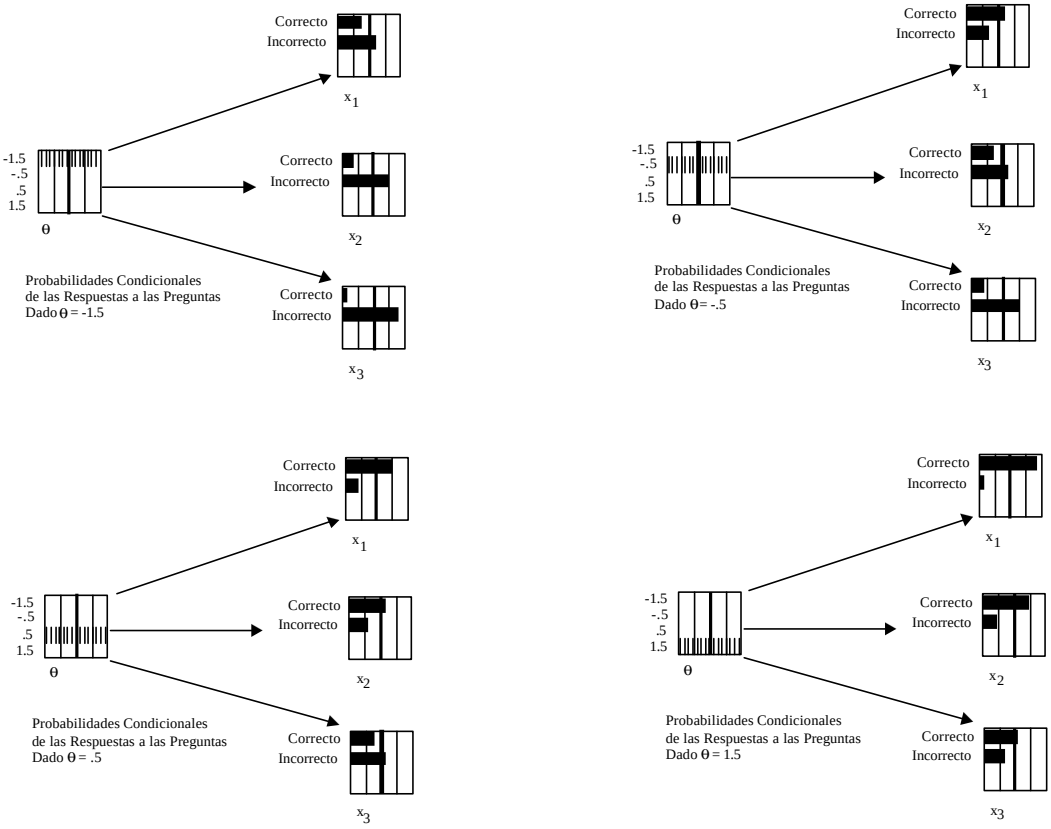
- Después de observar el dato x, la creencia de éste sobre el mismo conjunto de valores, se expresa como otra distribución de probabilidades (densidades), la distribución a posteriori $p(z \mid x)$.

- El valor evidencial del dato x es transmitido por la constante normalizadora, denominada **función de verosimilitud**: $p(x \mid z)$, que actualiza los valores de la distribución a priori con los valores de la distribución a posteriori, para todos los posibles valores de z. Una vez examinada la dirección en que cambian las creencias asociadas con una valor dado de z, en respuesta a la observación de la respuesta a la pregunta x, también cambia la intensidad con la que dichas creencias varían.

Tabla 1. Probabilidades de responder correctamente en el ejemplo IRT.

	Dificultad de la pregunta b		
Conocimiento del alumno q	-1	0	1
-1.5	.378	.182	.076
-.5	.622	.378	.182
.5	.818	.622	.378
1.5	.924	.818	.622

Estas relaciones se muestran como GADs en los 4 gráficos de la figura 3. Cada gráfico muestra las probabilidades de responder correcta e incorrectamente a cada pregunta si se sabe que θ toma una de los valores posibles. Las barras en los nodos que corresponden a preguntas representan las probabilidades de la tabla 1 para las respuestas correctas e incorrectas, dado los valores de θ. La barra para el nodo que representa a θ siempre vale 1 para el valor de θ fijado en cada caso (es decir, suponemos que el valor que toma θ en cada caso, tiene probabilidad 1 de ocurrir) y por esta razón, se actualizan las probabilidades condicionadas de las preguntas - $p(X_j = Correcta \mid \theta)$ - (deductivamente) como si dicho valor fuese el verdadero valor de θ.



Los nodos representan las variables; las barras representas las probabilidades de los posibles valores de una variable, sumando uno, con una barra rallada hasta uno representado la certidumbre

Figura 3. Probabilidades de las respuestas a la pregunta según el nivel de conocimiento del alumno.

4.3.3. IRT COMO UN MODELO GRÁFICO

La inferencia basada en probabilidades en redes complejas de variables interdependientes es una línea activa en investigación estadística, aplicable a tan

diversas aplicaciones como predicción meteorológica, análisis genealógicos, resolución de problemas y diagnóstico médico. La estructura de las relaciones entre las variables puede ser representada por un grafo acíclico dirigido (llamado comúnmente GAD), en el cual los nodos representan las variables y los arcos representan las relaciones de dependencia condicional. En correspondencia con el GAD existe una representación recursiva de la distribución conjunta de las variables de interés, denotada generalmente como $\{Z_1,\ldots,Z_m\}$:

$$p(Z_1,\ldots,Z_m)=\prod_{j=1}^{m} p\left(Z_j \mid \{\text{"padres" de } Z_j\}\right), \qquad (5)$$

donde los $\{\text{"padres" de } Z_j\}$ es el subconjunto de $\{Z_{j-1},\ldots,Z_1\}$ del que Z_j es directamente dependiente. En aplicaciones educativas, por ejemplo, se sitúan variables no observables que caracterizan aspectos del conocimiento del alumno como padres de las variables observables que caracterizan lo que los alumnos dicen y hacen en situaciones de evaluación.

La figura 4 muestra el GAD que corresponde al método IRT. Las variables genéricas Z se particularizan como la variable de conocimiento $\boldsymbol{q}$ y las variables de respuestas a las preguntas $\{X_1,\ldots,X_n\}$. El primer gráfico no muestra la dependencia de los parámetros de las preguntas, mientras que el segundo hace explícita la dependencia, indicando que la distribución de probabilidad condicional de cada X_j, dado $\boldsymbol{q}$, es una función de $\boldsymbol{b}_j$. Dicha estructura, que establece la independencia de las respuestas a una pregunta dada una sola variable no observable, es llamada a veces un modelo “casi Bayesiano” dado que raramente captura las sutiles relaciones encontradas en los problemas del mundo real. Aun así, este término despreciativo no es abandonado del todo en las implementaciones más meditadas del método IRT-CAT, ya que muchas variables que no aparecen en el modelo simple han sido manipuladas en un segundo plano expresamente para asegurar que su estructura simple sea suficiente para la tarea a realizar.

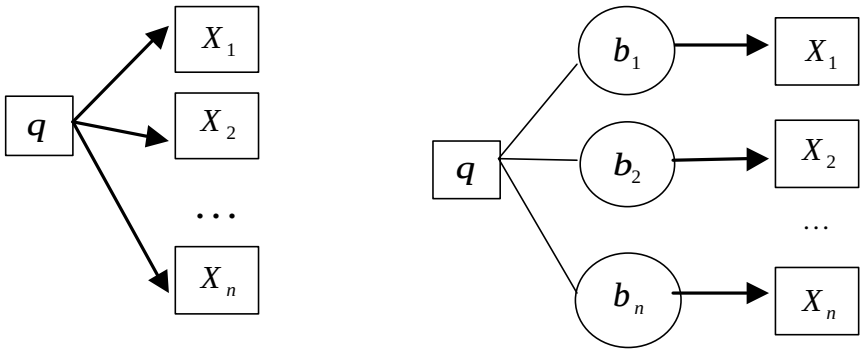


Figura 4. GADs para un modelo IRT. Los parámetros que determinan la distribución condicional de las Xs dado θ están implícitos en el gráfico de la izquierda y explícitos en el de la derecha.

Una forma de describir el método IRT-CAT desde la perspectiva del modelado gráfico es a través del GAD con θ como el único padre de todas las preguntas del banco de preguntas, tal y como se muestra en la figura 4. Al comienzo del test, la distribución marginal del nodo θ es $p(\theta)$. Cada pregunta es examinada hasta encontrar una que minimice la varianza esperada a posteriori; ésta es planteada al alumno, y el proceso se repite después de la respuesta de éste, comenzando ahora por $p(\theta \mid x^{(1)})$. El proceso continúa con cada $p(\theta \mid x^{(k)})$ sucesivo hasta que el test finalice. En cada paso, el valor observado de la variable administrada es fijo, la distribución de θ es actualizada, y los valores esperados para las preguntas aún no planteadas son revisados para calcular la varianza esperada a posteriori de θ si cada una de las preguntas fuera presentada al alumno la siguiente.

Una segunda aproximación para describir el método IRT-CAT es equivalente, estadísticamente hablando, pero subraya la modularidad del razonamiento que puede ser alcanzada con los modelos gráficos. La figura 5 muestra la situación en términos de fragmentos de modelos gráficos: la variable del modelo del alumno θ y una librería de nodos correspondientes a las preguntas del test, cualquiera de los cuales puede ser "unido" al nodo θ para producir un GAD, tal y como se muestra en el gráfico derecho de la figura. Este pequeño GAD es temporalmente unido a fin de recoger la evidencia sobre θ a partir de la respuesta dada a una pregunta *j* determinada. Es separado después de que la respuesta es observada y la distribución de θ es actualizada en consecuencia. El nuevo estado del conocimiento sobre θ proporciona una guía tanto para la selección de la siguiente pregunta a plantear, como para la terminación del test. Este proceso es un ejemplo de construcción de un modelo basado en el conocimiento.

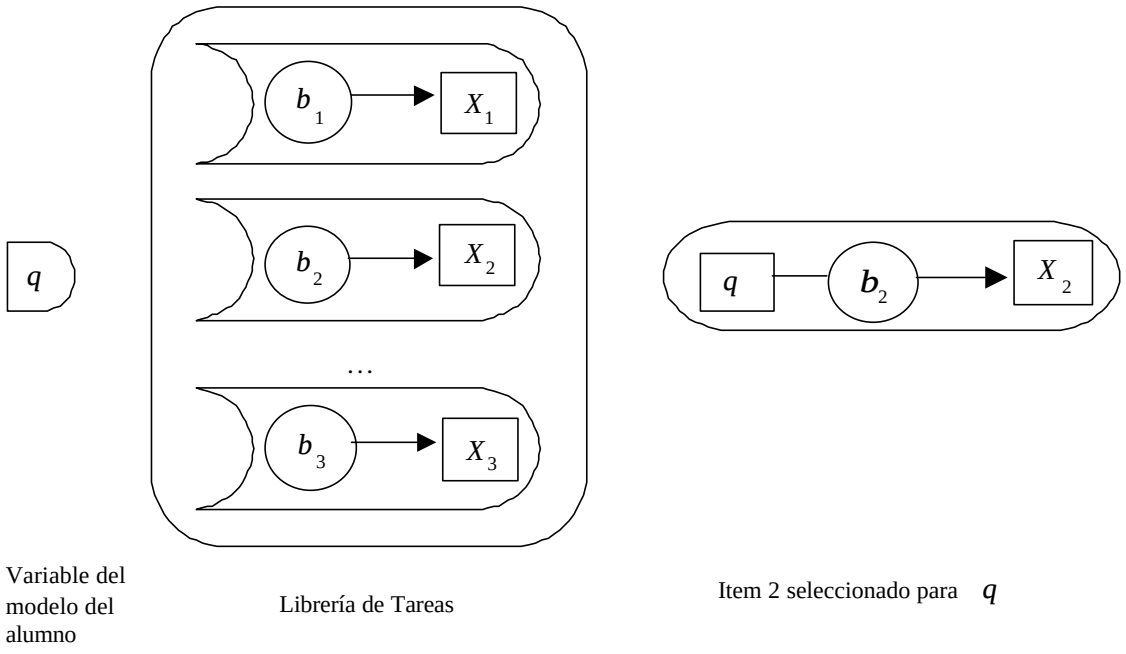


Figura 5. CAT como modelo de construcción basado en conocimiento. El gráfico de la derecha muestra el Item 2 unido al nodo q para crear un GAD dinámico.

4.3.4. EL PAPEL DE LAS VARIABLES EN EL MÉTODO IRT-CAT

En un primer acercamiento a los modelos IRT utilizados en los sistemas actuales, éstos dan la impresión de que todo lo que está ocurriendo puede ser entendido como modelos de alumno simples de una sola variable (el conocimiento total en cada área puntuable) y sus correspondientes bancos de tareas. Pero hay muchas más variables que están siendo manejadas de forma no aparente, algunas para definir efectivamente la variable que está siendo medida y otras, para asegurar que el modelo analítico simple caracterice de forma adecuada la información que está siendo adquirida.

Todos los problemas del mundo real tienen una mezcla única de características y necesidades, y cada persona tiene una aproximación única a sus necesidades. Esto es particularmente cierto para las tareas de evaluación, y por tanto, los examinandos variaran su grado de éxito en cada una de ellas. Las medidas educativas y psicológicas, tal y como han evolucionado durante este siglo, definen dominios de tareas de tal forma que las diferencias entre los examinandos con respecto a determinadas características tienden a acumularse a través de las tareas, mientras las diferencias respecto a otras tareas no tienden a acumularse. La variación que se acumula se convierte en “lo que

mide el test". Otras fuentes de variación generan incertidumbre sobre el nivel de un examinando en dicho test.

¿Qué prácticas han evolucionado para guiar la construcción de tests bajo esta perspectiva?. Esta sección discute el papel que estas variables tienen para dicho fin en los sistemas IRT-CAT:

- Estas variables pueden limitar el ámbito de la evaluación, y no aparecer nunca en el modelo analítico.
- Estas variables pueden describir características de las tareas, a fin de construir tareas y modelar los parámetros de las preguntas.
- Estas variables pueden controlar la construcción del test.
- Estas variables pueden caracterizar a las respuestas (variables observables)
- Estas variables pueden caracterizar determinados aspectos del conocimiento del alumno(colectivamente, constituyen el modelo del alumno)

Una variable dada, puede jugar papeles diferentes en tests diferentes, de acuerdo con los propósitos de estos tests. Sólo las variables que juegan el último papel en la lista anterior aparecen de forma explícita en el modelo (en el caso de IRT-CAT, q). q es tomada de forma práctica como un resumen de la evidencia acerca de una construcción conseguida a través de elecciones hechas sobre, y manipulación de, muchas otras variables "ocultas" en los cuatro primeros puntos de la lista anterior.

4.3.4.1. Variables que limitan el ámbito de la evaluación

Esta sección muestra como dos tipos de estudios cuyo propósito inicial es el análisis de la validez, ayudan a asegurar que la simple estructura de IRT es la adecuada. En ambos casos, las variables que pueden generar interacciones entre las respuestas a las preguntas, más allá de las que están representadas por una variable global del nivel de conocimiento, son el punto central del estudio, y se toman determinadas acciones de forma que dichas variables no necesiten ser incluidas en el modelo analítico. Los resultados del primer estudio llevan a restringir los contextos de los tests y de los métodos, de forma que la variable operativa definida q esté condicionada de forma efectiva por valores específicos de dichas variables. Los resultados del segundo estudio pueden llevar a eliminar preguntas que puedan engendrar fuertes interacciones con

características del alumno no modeladas, de forma que se puedan despreciar dichas variables sin que afecten al modelo.

Delimitar el dominio y los métodos de evaluación. Multitud de aspectos como el nivel, conocimiento y experiencia del alumno, afectan a su rendimiento en cualquier dominio de aprendizaje, el cual no puede, ni debe, ser totalmente abarcado por ningún test en concreto. Debemos considerar qué aspectos del universo de potenciales tareas de evaluación, son los más destacados para el trabajo a realizar y, determinar cuales de ellos incluir en el test y cuales excluir de él. El camino que elijamos para evaluar el conocimiento en los tests, tiene un efecto significativo en la evaluación; algunos examinandos son relativamente mejores frente a un tipo de tareas que frente a otras, obtienen mejores resultados en algunos tipos de tests que otros, o están más familiarizados con unos contextos que con otros. Habrá, por tanto, tendencia a tener unas asociaciones más fuertes que otras relativas al contexto y a los métodos del test – interacciones que invalidan la estructura del GAD de la Figura 1.

Funcionamiento Diferencial de las Preguntas (DIF; Differential Item Functioning). El funcionamiento diferencial de las Preguntas (DIF) ocurre cuando, por razones no relacionadas con el nivel o el conocimiento de interés, el contenido de ciertas tareas o las características del formato tienden a ser relativamente duras para los miembros de una subpoblación identificada, caracterizadas, por ejemplo, por un determinado genero (masculino o femenino) o procedencia étnica. El GAD de la figura 6 muestra esta situación inesperada. Los análisis DIF exploran los datos anteriores al test para detectar su presencia. Así, una aplicación de enseñanza puede buscar a propósito preguntas para las cuales el interés personal es alto para ciertos estudiantes, de forma que se motive mejor a éstos para que comprendan los conceptos a los que se refieren.

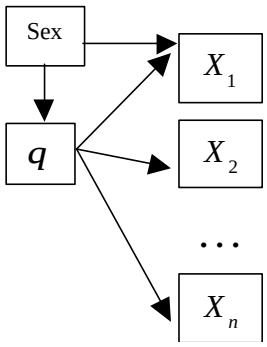


Figura 6. Un GAD ilustrando DIF. Las probabilidades de las respuestas de los ítems 2 a n son condicionalmente independientes del sexo, dado q. Las probabilidades para la respuesta al ítem 1 son dependientes tanto del sexo como de q.

4.3.4.2. Variables que definen características de las tareas

Las tareas individuales en un test pueden ser descritas en términos de múltiples variables. Abarcan cosas tales como el formato, contenido, modalidad situación, propósito, carga de vocabulario, estructura gramatical, conocimientos matemáticos requeridos, etc. Algunas de estas variables aparecen formalmente en las especificaciones del test, pero los desarrolladores de tests emplean muchas más durante la creación de las tareas. Sin codificar o nombrar formalmente esta información en términos de variables, los escritores de tareas trabajan con fuentes como resultados anteriores con preguntas similares, experiencia acerca de cómo los estudiantes aprenden los conceptos, conocimiento de las equivocaciones más comunes e investigación cognitiva sobre el aprendizaje y resolución de problemas en el dominio. Los estudios han mostrado que estos tipos de variables pueden dar una predicción muy fuerte de la dificultad de una pregunta.

Una forma de utilizar esta información colateral sobre las tareas puede ser complementar, quizás suplantar, los datos provenientes de muestras de los examinandos antes del test como la fuente de información acerca de los parámetros B de las preguntas en el modelo IRT. En efecto, se puede crear un GAD de segundo orden para modelar los parámetros de las preguntas (figura 7).

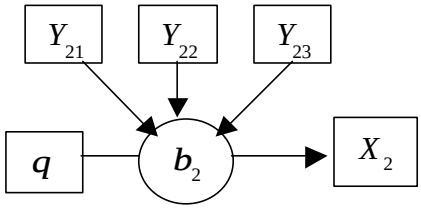


Figura 7. Una parte de un GAD de dos niveles.

Una segunda forma de utilizar las variables normalmente ocultas que caracterizan las preguntas del test puede ser crear una red más basada en estos principios, a la hora de construir las preguntas. Dichas variables pueden ser la base de "esquemas de preguntas" o "interfaces de preguntas", para el desarrollo de familias de tareas con características y propiedades que son a la vez bien comprendidas, y basadas

de forma demostrable en un marco teórico del conocimiento y del nivel, que el test debe determinar. Las características de esquemas y las características de los elementos que llenan los esquemas pueden ser utilizadas entonces para modelar los parámetros IRT, tal y como se comentaba anteriormente.

Una tercera vía para utilizar las variables que caracterizan los requisitos de las tareas es unir los valores de las variables del modelo del estudiante con los comportamientos observables esperados. Con el modelo de Rasch, por ejemplo, el conocer b_j nos permite calcular la probabilidad de una respuesta correcta de parte de un estudiante que tiene un valor q determinado. Recíprocamente, podemos dar un significado a un valor de q describiendo el tipo de preguntas a las que es posible que responda correctamente un estudiante de ese nivel y el tipo de las que no. A la larga, estas características de las preguntas son proporcionadas por bs y por lo tanto podremos describir el nivel de conocimiento de un alumno en términos de las características de las preguntas y/o de los niveles cognitivos relevantes.

4.3.4.3. Variables que controlan la creación del test

Una vez se ha determinado un dominio de preguntas, las especificaciones del test restringen la mezcla de preguntas que constituyen el test de un alumno dado. Nosotros no observaremos ni el total del dominio de la tarea ni una muestra incontrolada, sino una composición construida cuidadosamente bajo reglas preespecificadas para el "bloqueo" y el "solapamiento".

Las restricciones de bloqueo aseguran que aunque a diferentes examinandos se les presenten diferentes preguntas, generalmente de dificultad diferente en un CAT, nunca obtendrán una mezcla similar de contenido, formato, modalidades, demandas de nivel, etc.

Las restricciones de solapamiento abarcan las innumerables características de idiosincrasia de las preguntas que no pueden ser codificadas y catalogadas de forma exhaustiva. Se especifican conjuntos de preguntas que no pueden aparecer en el mismo test debido a que se dan respuesta unas a otras o examinan del mismo concepto. Las restricciones de solapamiento han evolucionado a través de líneas sustantivas mas que de líneas estadísticas, a partir de la intuición de que las preguntas solapadas reducen la información sobre los examinandos. El formalismo del modelado gráfico nos permite

explicar por qué, cómo y cuanta información se pierde. Cada pregunta es aceptable por si misma, pero su aparición conjunta puede introducir una fuerte dependencia condicional inaceptable.

La figura 8 ilustra el impacto de las restricciones en la creación de tests con un ejemplo simple. El banco de preguntas contiene sólo cuatro preguntas; las preguntas 1 y 2 utilizan ambas la palabra “ubicuo” que no es familiar, y las preguntas 3 y 4 tratan las dos sobre los triángulos rectángulos. Las restricciones de solapamiento pueden decir que una pregunta de cada par debe aparecer en el test de cada examinando. Los gráficos primero y segundo de la figura 8, son GADs alternativos para el banco completo, uno mostrando las dependencias condicionales entre los conjuntos solapados y el otro introduciendo variables adicionales del modelo del alumno. El tercer gráfico es el GAD estándar de IRT-CAT con las restricciones de solapamiento y bloqueo colocadas en su sitio –su simplicidad es apropiada porque el flujo entrante de evidencia ha sido restringido de forma que evite algunas violaciones particulares de su fuerte estructura de independencia condicional.

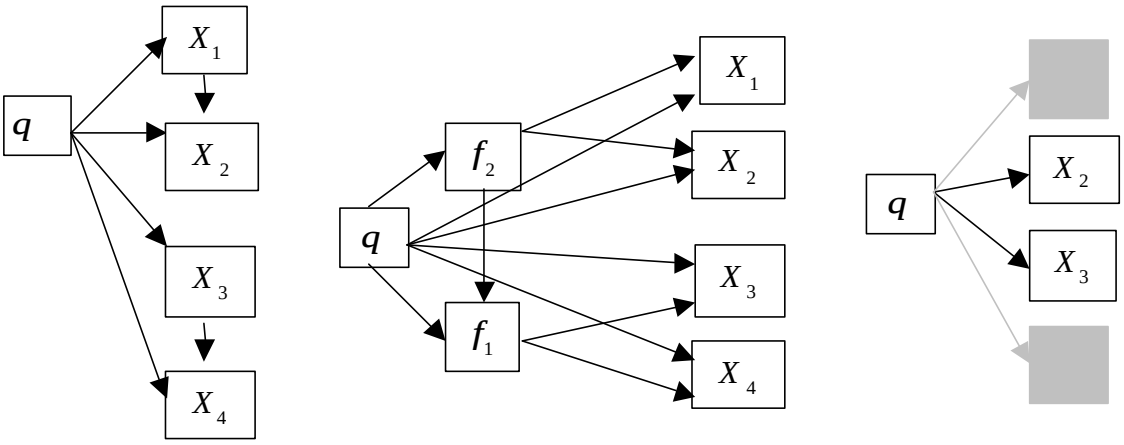


Figura 8. Tres GAD relacionados con las restricciones de solapamiento. El primer gráfico muestra la dependencia condicional entre conjuntos de ítems. El segundo, muestra la independencia condicional a la que se llega si se añaden variables al modelo del alumno. El tercer gráfico muestra la independencia condicional que se alcanza en el modelo IRT si se observan restricciones.

Muchas otras variables pueden ser definidas para caracterizar las preguntas de los tests de acuerdo con las características no controladas por las restricciones de bloqueo o solapamiento. Aquellas incluyen las variables a nivel de pregunta discutidas en la sección "Variables que describen las características de las tareas" que pueden ser

utilizadas para modelar los parámetros de las preguntas, así como muchas de las características incidentales o de idiosincrasia que hacen a cada pregunta única. Estas variables son tratadas utilizando la aleatoriedad; los valores particulares que toman en cualquier test de un alumno dado, son una muestra aleatoria del banco de preguntas, sujeta a las restricciones de bloqueo, solapamiento y medida.

4.3.4.4. Variables que caracterizan las respuestas (observables)

La caracterización de las respuestas de los estudiantes está directamente relacionada con las preguntas de múltiple elecciones en IRT-CAT: ¿Indicó el estudiante la opción preespecificada como correcta u otra diferente? Las respuestas abiertas pueden ser analizadas también utilizando los modelos dicotómicos IRT, pero se necesita mas juicio para obtener la "corrección" para generar notas únicas. En estos últimos casos, las variables pueden ser definidas para describir calidades de los productos o los productos de las notas de los estudiantes, y las reglas pueden ser revisadas para mapear los valores de estas variables en la dicotomía correcto/incorrecto.

De forma más general, las características salientes de las respuestas de los examinandos pueden ser codificadas en términos de categorías de calificación total o parcialmente ordenadas. Los modelos IRT han sido extendidos más allá de los datos dicotómicos para tratar con estas categorías ordenadas de respuestas. En este caso, modelar X_j puede ser tan inmediato debido a las restricciones sobre el posible comportamiento de la respuesta, o bien, puede necesitar un paso adicional de evaluación en términos de las propiedades abstraídas de los comportamientos menos restringidos de las respuestas.

4.3.4.5. Variables que caracterizan aspectos del nivel de conocimiento (el modelo del alumno).

Las variables de los modelos del alumno integran información a lo largo de distintas partes de evidencia para soportar la inferencia sobre el nivel y el conocimiento de los estudiantes a un nivel de abstracción mayor que el particular de cada una de las tareas específicas (un nivel acorde con la instrucción, documentación y toma de decisiones como demanda la aplicación). La naturaleza de las variables de los modelos del alumno debe ser dada por el propósito del test, pero también debe ser consistente

con los patrones empíricos de respuesta y con las teorías de la capacidad dentro del dominio. Como se discute en las secciones siguientes, no es posible ni deseable el incluir en las variables del modelo todos los aspectos concebibles de capacidad del alumno. La elección viene determinada por los propósitos utilitarios, tales como las distinciones que serán importantes para la toma de decisiones.

Por ejemplo, el sistema TOEFL (sistema para la evaluación del lenguaje) actual tiene tres variables en el modelo del estudiante –entendimiento oral, lectura y estructura gramática, o L, R y S- y cada una es evidenciada por tareas discretas que son de su mismo tipo, con dominios disjuntos de preguntas y las variables del nivel de conocimiento asociadas en el dominio, q_L, q_R y q_S. Estas variables son utilizadas para decisiones no muy frecuentes pero trascendentes tales como la admisión de hablantes ingleses no nativos en los programas para graduados. En contraste, un sistema tutorial inteligente (STI) debe definir variables de modelos de estudiante con una granularidad más fina de forma que puedan proporcionar de forma frecuente y específica la educación a los alumnos. El principio rector de los STIs es que los modelos del alumno deben ser especificados al mismo nivel en el que son tomadas las decisiones educativas.

Los sistemas IRT-CAT estándar están basados en modelos de estudiante univariantes. Los modelos de estudiante multivariantes cobran mayor importancia cuando las observaciones contienen información acerca de más de un aspecto del conocimiento, para los cuales es deseable el acumular evidencias.

4.3.5. GENERALIZACIÓN MULTIVARIABLE DEL MODELO DEL ALUMNO.

Los expertos difieren de los novatos, no únicamente por dominar más hechos y conceptos, sino también por tener presente y aprovechar las interconexiones entre ellos. La valoración directa de la experiencia creciente requiere, además,

♦ tareas más complejas para poder extraer la evidencia que se esconde tras los aspectos múltiples e interrelacionados del nivel de conocimiento, y

♦ modelos de estudiante multivariantes, de forma que capturen, integren y acumulen la importancia de las prestaciones de los estudiantes a través de dichas tareas.

El hecho de que los modelos IRT estándar no llegan al nivel exigido, no hace que sea necesario abandonar sus principios básicos de inferencia, sino que es necesario más bien extenderlos. Podemos construir sobre las mismas, ideas de definir variables no observables para "explicar" los patrones observados de las respuestas, y que "algunas formas de variación se acumulan y otras no" –y la utilización de inferencia basada en probabilidades para controlar el conocimiento acumulado y la incertidumbre restante sobre el conocimiento del alumno cuando la valoración se esté realizando. Esta sección esquematiza una aproximación en términos generales, resaltando como resuelve ésta la problemática discutida anteriormente en el contexto de los sistemas IRT-CAT.

La figura 9 ilustra una posible implementación de un sistema de evaluación adaptativa basada en modelos gráficos (GM-CAT). Se presenta aquí como una referencia visual para la discusión de las propiedades matemáticas. (Debe ser notado que algunas de las variables –en particular, q_R, q_W, q_S y q_L- evocan los conceptos Lectura, Escritura, Habla y Escucha. Sea cual sea el significado que se pretende al introducir estas variables en el modelo, su significado operacional es una media de la puntuación en tareas relacionadas con estas modalidades. De todas formas, el significado real de las variables del modelo está controlado por variables que no aparecen para nada en la figura 9: esto es, las variables que controlan el ámbito del test [Sección "Variables que limitan el ámbito de la evaluación"] y la selección de tareas [Sección "Variables que controlan la creación del test"]).

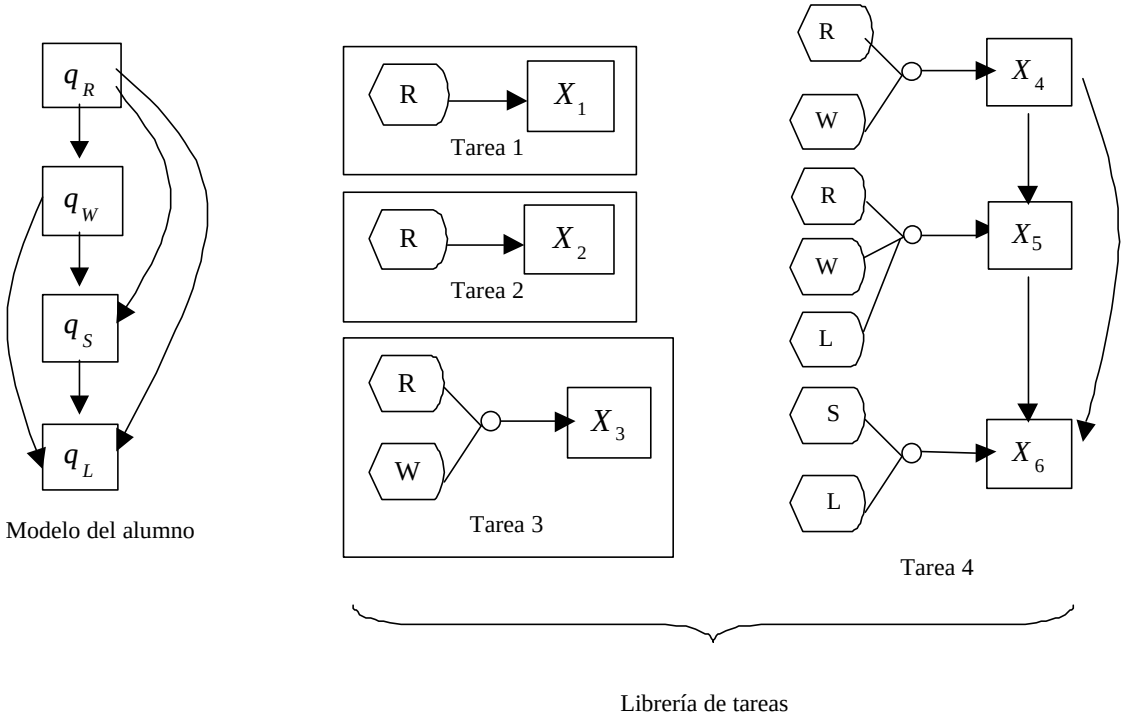


Figura 9. Un GAD orientado a tareas.

El modelo en el marco GM-CAT está distribuido entre dos fuentes. A la izquierda está el modelo del alumno, el cual es fijo a lo largo de todas las administraciones de los tests. A la derecha, está la colección de los modelos de las tareas/evidencias, o fragmentos de GAD, que corresponden a un banco de preguntas. Un examinando dado, verá un subconjunto de las tareas de acuerdo con el *algoritmo de selección de tareas*, el cual hace un balance entre las consideraciones de la información y las restricciones de contenido y solapamiento. Cuando una tarea es asignada a un examinando, el modelo de evidencia asociado con dicha tarea es "enganchado" con el modelo del estudiante (de acuerdo con el patrón de unión de variables del modelo de evidencias). La evidencia de la respuesta del examinando a la tarea es recogida entonces por el modelo principal del alumno, y en el modelo de tarea/evidencia puede ser "desenganchado", dejando el modelo del alumno actualizado para la siguiente tarea. De esta forma, el marco GM-CAT es otra aplicación de construcción de modelos basados en el conocimiento.

Los nodos del modelo del alumno son variables no observables relacionadas con el rendimiento del alumno. –una generalización multivariable del papel de q en IRT. Las variables del modelo del alumno representan aspectos del nivel de conocimiento y son incluidas en el modelo tanto para ser utilizadas para reportar el rendimiento del

alumno, como para acumular patrones suplementarios en tareas con propósito de diagnóstico, como para tomar en cuenta dependencias incidentales entre tareas. Su naturaleza y número deben ser consistentes con, pero no determinadas únicamente por, un conocimiento del nivel del alumno en el dominio. La determinación final del número y granularidad de las variables pertenecientes al modelo del alumno está gobernado por los requisitos de generación de informes y diagnóstico de la evaluación. De esta forma un test de licenciatura que puede conllevar el pase del alumno a otro curso o su repetición debe utilizar un modelo del alumno mucho más grueso que el de un sistema tutorial inteligente.

Los nodos en los modelos de evidencia de las tareas son variables observables que corresponden a aspectos sobresalientes del comportamiento de los examinandos en situaciones de tareas específicas (una generalización de las respuestas a las preguntas de IRT). Generalmente, estas corresponderán a características de la respuesta a la tarea. Pueden ser tan simples como "¿dio el examinando una respuesta correcta a una pregunta de múltiples respuestas?" o tan complejas como las dimensiones de una tabla de puntuación de múltiples atributos producida por un juez humano.

Existen tres tipos de asociaciones entre el modelo del estudiante y los nodos observables:

$\Rightarrow$ El primer tipo de asociación es la más importante: Las variables del modelo del estudiante son padres de variables observables. De esta forma, el nivel y el conocimiento "explican" los patrones en el comportamiento observable en las tareas a mano, y cuando son observadas las respuestas, el grado de credibilidad sobre las variables del modelo del alumno se actualiza. Las asociaciones toman la forma de probabilidades condicionales de los valores de las variables observables, dados los valores de las variables del modelo del estudiante (una generalización de los parámetros de las respuestas en IRT). Cuando aspectos múltiples del nivel y el conocimiento son situados como padres de una variable observable dada, las relaciones tales como conjunción, disyunción y compensación pueden ser propuestas. Los diseñadores de las tareas indican la estructura de estas asociaciones (indicadas por las uniones de las preguntas en la figura 9) y proporcionan las estimaciones iniciales de las probabilidades condicionales basadas en las variables de las características de las tareas, variables de características de las respuestas, y las expectativas de éstas últimas dadas las primeras en varios niveles de las variables del modelo del alumno. Estas probabilidades

condicionales pueden ser modeladas más allá como funciones de las variables características, como una generalización de la técnica IRT descrita en la figura 7.

$\Rightarrow$ Un segundo tipo de asociación es aquella entre variables observables, sobre y bajo las asociaciones inducidas por las variables del modelo del alumno. Estas ocurren cuando son capturados como observables múltiples aspectos de la prestación en una misma situación de tarea, e incluidos en el GAD como un medio de modelar los efectos de los contextos compartidos, las similitudes en los métodos de respuesta, o conexiones incidentales que las restricciones de solapamiento deben prohibir en IRT-CAT. Un modelo de tarea/evidencia para una tarea compleja debe comprender múltiples variables observables, quizás con asociaciones engendradas por las partes comunes inducidas por el contexto compartido, pero probablemente con diferentes padres del modelo del alumno, de acuerdo con sus demandas particulares. Estas asociaciones son ilustradas en la figura 9 por las flechas que conectan las variables observables X_4, X_5 y X_6.

$\Rightarrow$ Un tercer tipo de asociación es aquella que relaciona varias variables del modelo del alumno: esto es, algunas variables del modelo del alumno pueden aparecer como padres de otras variables del modelo del alumno de forma que se expresen dichas relaciones como prerrequisitos, correlación empírica o relaciones lógicas tales como la conjunción y la disyunción. Estas asociaciones aparecen en la figura 9 como flechas que conectan variables del modelo del alumno con otras variables del modelo del alumno. De esta forma, la evidencia directa sobre una variables del modelo del alumno puede proporcionar una evidencia indirecta acerca de otra, asimismo se pueden explotar las asociaciones entre niveles o capacidades para mejorar la precisión de los informes generados.

La evaluación adaptativa con un modelo gráfico puede utilizar el estado actual del modelo del estudiante como parte del algoritmo de selección de la siguiente pregunta. De la misma forma que IRT-CAT, el método GM-CAT selecciona las tareas de un banco de tareas de forma que se maximice alguna medida de información. El *valor de información* y el *peso de la evidencia* parecen candidatos prometedores. El sistema GM-CAT une el modelo de la evidencia/tarea al modelo del alumno y recoge la evidencia proporcionada por la respuesta del examinando. El algoritmo puede entonces descartar la tarea o mantenerla en el modelo si es necesario para tratar los efectos de dependencia entre tareas (en efecto, consideraciones de solapamiento tratadas por el modelo, en lugar de evitadas). El algoritmo todavía va a necesitar balancear los

contextos de las tareas, el contenido, los tipos de tareas y demás entre examinandos, dado que estas especificaciones definen operacionalmente las variables del modelo del alumno en el mismo sentido que los bancos de preguntas y la construcción del test definen el parámetro q en IRT:

El estado del modelo del alumno es utilizado también para la generación de informes, o, en aplicaciones interactivas, para disparar la realimentación. Si se desea un resumen en formato de número simple del rendimiento del alumno, uno puede proyectar el estado actual del modelo del alumno sobre una dimensión particular. Los estudios de validez crecen en importancia, a causa de que la validez del modelo interno debe ser monitorizada ahora, así como las relaciones con las variables externas al modelo.

4.3.6. SIGUIENTES PASOS

Un entendimiento claro de que es lo que está implicado en las aplicaciones exitosas de los sistemas IRT-CAT es un primer paso útil hacia la extensión de la aproximación a sistemas más complejos. La inferencia basada en probabilidades con los modelos gráficos ofrece un marco para expresar y más tarde afrontar, los problemas que puedan surgir. Sin tener en cuenta los éxitos preliminares, existen aun un gran número de asuntos que deben ser resueltos para desarrollar una teoría de la valoración basada en modelos gráficos, que abarque tanto la evaluación como los sistemas CAT. Hemos mencionado anteriormente la importancia de los fundamentos cognitivos de una aplicación. Entre los desafíos técnicos que hemos comenzado a hacer frente están:

Construcción de modelo basado en conocimiento (KBMC; Knowledge-Based Model Construction). KBMC está relacionada con la construcción dinámica y la manipulación de modelos gráficos, adaptándose a los cambios en el estado de conocimiento pero en relación con la importancia de las preguntas que están siendo planteadas; esto es, revisando los modelos para reflejar marcos cambiantes de discernimiento, así como los estados cambiantes de conocimiento y las cambiantes situaciones externas. Los sistemas IRT-CAT se adaptan a los cambios de estado de conocimiento dentro de un marco estático de discernimiento –la pregunta es siempre "¿Qué es $\boldsymbol{q}$?"- y utiliza formulas de información, bloqueo basado en tareas y restricciones de solapamiento para seleccionar preguntas. Es necesaria la generalización de estas reglas para poder afrontar modelos más complejos en los cuales las diferentes subpartes del modelo pueden entrar y salir del campo de atención.

Dependencias inducidas por la tarea. Un modelo evidencia/tarea puede proporcionar descendientes comunes de dos variables condicionalmente independientes en el modelo del estudiante. El colapsado de tareas producirá nuevos arcos en el modelo del estudiante. La teoría GM-CAT requerirá tanto técnicas de aproximación para la determinación de dichos arcos cuando pueden ser observados, como técnicas para la recompilación dinámica del árbol de uniones.

Variables continuas en el modelo del estudiante. El modelo gráfico más común que contiene tanto variables continuas como discretas es el modelo Condicional Gausiano (CG). Todos estos modelos tienen variables continuas (normales) condicionadas por las variables discretas. En la evaluación educacional, sin embargo, parece más natural tener como variables discretas las respuestas a las preguntas, condicionadas éstas por los continuos rendimientos del estudiante. Quizá el IRT (una extensión multivariable del modelo de Rasch) pueda entrar en servicio aquí, pero la falta de una solución cerrada requerirá soluciones numéricas que pueden entrar en los requisitos dinámicos de CAT. La dificultad estriba en que no existen soluciones cerradas cuando las variables continuas son padres de las preguntas discretas.

Ajustado en el modelo. Modelos de alumnos más complejos y variables de rendimiento de la tarea, incrementan el esfuerzo del analista para ajustar, probar y mejorar los modelos. Una ventaja particular de la utilización de la inferencia basada en probabilidades es que se pueden utilizar técnicas estadísticas estándar para abordar muchas de estas cuestiones, en conexión con el uso de redes bayesianas en sistemas expertos de forma más general. Además se puedan adaptar diagnósticos más especializados para los modelos con variables no observables provenientes de la literatura psicométrica.

4.4. PROCEDIMIENTOS PARA SELECCIONAR LAS PREGUNTAS EN LOS TESTS ADAPTATIVOS ASISTIDOS POR ORDENADOR.

Los procedimientos estándar para seleccionar preguntas pueden ser clasificados en dos grandes grupos:

- pre-estructurados
- de intervalo variable

La mayor diferencia entre ellos está en la cantidad de adaptabilidad que permiten los procedimientos. La selección pre-estructurada da como resultado búsquedas deterministas, casi fijas, a través del test. La selección de intervalo variable permite ramas interactivas o caminos ilimitados a través del test.

4.4.1. SELECCIÓN DE PREGUNTAS PRE-ESTRUCTURADA.

El intento inicial en los tests adaptativos es usar bancos de preguntas pre-estructurados para definir un algoritmo de selección. Cuatro de los procedimientos más conocidos en los tests adaptativos son:

- "en dos etapas"
- "piramidal"
- "de nivel flexible"
- "adaptativo estratificado"

Las técnicas de selección pre-estructurada no son las más adecuadas para un test adaptativo porque:

- limitan la cantidad de elementos entre los que realizar la selección y,
- no utilizan toda la información disponible sobre un alumno.

4.4.1.1. Selección Adaptativa Estratificada

Con esta técnica, el banco de posibles preguntas es dividido en un número de estratos (9 es el número comúnmente usado) según la dificultad de las preguntas que almacenan. De este modo, el primer estrato debería contener las preguntas disponibles de menor dificultad, mientras que el mayor estrado debería contener las preguntas más difíciles. Dentro de cada estrato, las preguntas serán ordenadas de acuerdo a sus índices de discriminación (suponiendo que se usa un modelo IRT de 2 ó 3 parámetros). Un alumno podría comenzar el test desde cualquier estrato, dependiendo de la información disponible a priori sobre su nivel de conocimiento. Además, todos los alumnos podrían empezar el test dentro del mismo estrato, si no existe información a priori sobre ellos.

El test comienza con la pregunta que mayor índice de discriminación posea dentro del estrato elegido. Si la persona responde correctamente a la pregunta, la siguiente pregunta será aquella con el mayor índice de discriminación en el estrato que

sigue al actual al incrementar la dificultad. El test continúa de este modo: subir un estrato si la respuesta es correcta y bajar un estrato si la respuesta es incorrecta.

El procedimiento siempre presenta las preguntas con mayor índice de discriminación del estrato apropiado, que aún no se han presentado. El test acaba cuando se satisfaga el criterio de finalización fijado (es decir, que el número de preguntas planteadas alcance un determinado valor, etc.).

Esta estrategia de selección adaptativa tiene importantes características que la resaltan frente al resto de procedimientos pre-estructurados, haciendo que esta técnica aún sea usada. El procedimiento permite variar el nivel de entrada al test, haciendo posible que el procedimiento de selección se ajuste al nivel de dificultad adecuado más rápidamente. Además, con este procedimiento se produce una gran diferencia (intervalo) en la dificultad de dos preguntas planteadas consecutivamente. Este gran intervalo permite a los alumnos moverse desde preguntas de dificultad media a las preguntas de mayor dificultad con tan sólo unas pocas de respuestas correctas consecutivas. Este hecho enfatiza la eficiencia del test.

Desafortunadamente, el intervalo utilizado en esta estrategia permanece constante a lo largo de todo el test, por lo que un intervalo grande podría ser considerado una desventaja de la estrategia a la hora de finalizar el test, ya que el nivel de dificultad de las preguntas debería estar cerca del nivel de conocimiento del alumno. Si el nivel de conocimiento del alumno permanece cerca del nivel de dificultad de la pregunta de un estrato, el alumno puede realizar, al menos, la mitad de las preguntas pertenecientes a ese estrato (debido a que el método cambia de estrato después de la respuesta a cada pregunta). Así, al menos la mitad de las preguntas en el test proporcionarían una medida menor a la óptima para ese alumno.

La eficiencia de esta estrategia podría mejorarse si se consigue plantear preguntas a los alumnos, de un nivel de dificultad más apropiado. Dicha mejora podría consistir, en usar un gran intervalo al principio del test, cuando aún se tiene poca información sobre el alumno (se habilita así una rápida progresión al nivel de dificultad apropiado para dicho alumno). Este intervalo debería disminuirse a medida que el test progresa, reflejando la estimación actual del conocimiento del alumno. Aquellos procedimientos que varían el intervalo de dificultad en función del grado de confianza en la puntuación actual del test, son los más eficientes. Dos de estos procedimientos se discuten en el apartado siguiente.

4.4.2. PROCEDIMIENTOS DE INTERVALO VARIABLE.

4.4.2.1. Método de Máxima Información.

El procedimiento de selección de la máxima información es conceptualmente simple aunque proceduralmente complejo. Este método elige aquella pregunta que maximiza la información cuando el alumno la responde. Para seleccionar la mejor pregunta, se calcula la información proporcionada por cada pregunta existente en el banco, dado la estimación actual del nivel de conocimiento del alumno y los parámetros de la pregunta. La pregunta que proporcione la máxima información sobre el nivel de conocimiento del alumno es la seleccionada. Todos estos cálculos, en la mayoría de los casos, se producen después de que el alumno responda a la pregunta en curso y antes de seleccionar la próxima pregunta.

Las ventajas proporcionadas por el método de la máxima información frente a los procedimientos pre-estructurados son considerables. En primer lugar, ya que la pregunta que proporciona la máxima información es la seleccionada cada vez, la eficiencia del test se incrementa. A su vez, la estimación del conocimiento se va haciendo mucho más precisa cada vez que se plantea una pregunta, ya que el intervalo inicial es bastante grande pero disminuye conforme el alumno responde a las preguntas, permitiéndose así, una rápida progresión hacia el rango correcto de dificultad de las preguntas a plantear. Además, este procedimiento permite muchos caminos a través de un test (debido a que se busca la mejor pregunta del banco cada vez que se estima un nuevo nivel de conocimiento para el alumno) proporcionando una medida de seguridad, ya que cada examinando recibirá diferentes preguntas.

Las desventajas de este método proceden de las mismas características que han sido expuestas como ventajas. Es decir, el proceso de búsqueda sobre la totalidad del banco para encontrar la mejor pregunta a plantear, conlleva un gran número de cálculos estadísticos que consumen un tiempo bastante grande, si el banco tiene un tamaño considerable. Por tanto, para bancos de preguntas muy grandes, este método (tal y como se ha descrito) causaría retardos a la hora de plantear las preguntas a los alumnos, cosa impracticable en la mayoría de situaciones. A su vez, la confianza en las estimaciones provisionales de los niveles de conocimiento, amplía la necesidad de la exactitud en los procedimientos que las calculan.

4.4.2.2. Método Bayesiano.

El método de selección Bayesiana (Owen, 1975) es similar al método de la máxima información. Cada alumno comienza el test con una estimación inicial de su nivel de conocimiento y con un intervalo de confianza asociado a dicha estimación. Estos datos son calculados como la media y la varianza, de una distribución normal a priori del conocimiento que está siendo medido. Cada vez que una pregunta es contestada, se calcula las nuevas estimaciones usando la respuesta dada a la pregunta y los valores de la distribución a priori, y con dichos valores se crea la nueva distribución a posteriori. El método de selección bayesiano elige aquella pregunta que más minimice la varianza a posteriori (es decir, se intenta reducir el error de la estimación al máximo). Específicamente, la varianza a posteriori se calcula para todos las preguntas del banco, dado el nivel de conocimiento inicial del alumno y los parámetros de la pregunta (nivel de dificultad, factor de adivinanza e índice de discriminación). Se elegirá aquella pregunta que disminuya la varianza a posteriori al menor valor.

Todas las ventajas y desventajas expuestas en el método de la máxima información se producen también con el método de selección bayesiano. Por otro lado, se piensa, que este método proporciona más estabilidad en el rango de dificultad e información que el método de la máxima información, pero esto aún no ha sido probado. Si el tamaño del banco de preguntas se incrementa, el tiempo empleado en buscar la pregunta óptima a plantear al alumno se incrementará también, ralentizando el proceso de generación del test.

Algunos autores han criticado el uso del método bayesiano como método de puntuación, sobre todo en tests de tipo educativo. El argumento que han dado es que dicho método se basa en la distribución inicial del conocimiento de cada individuo. Por este motivo, dos individuos que responden exactamente a las mismas preguntas pueden llegar a recibir puntuaciones diferentes si las estimaciones a priori de sus niveles de conocimiento son diferentes. De ahí, que dichos autores califiquen este método como método de puntuación apropiado a nivel de grupo (ya que reduce el error de la varianza en las puntuaciones) pero a nivel de individuo puede ser inapropiado.

Si en un test adaptativo se utiliza el método de selección bayesiano junto al método de puntuación de máxima verosimilitud, el algoritmo podría ser bastante eficiente.

Desde un punto de vista puramente estadístico, la generación de preguntas a plantear al alumno en un test adaptativo se convierte en un problema de cálculo de probabilidades (Owen, 1975). Por tanto, las respuestas que el alumno da a cada pregunta serán las observaciones X_1, X_2, X_3, ... que se producen secuencialmente en el tiempo y que son independientes. Cada observación, X_i, será un 0 o un 1 y la distribución de X_i estará especificada por:

$$P(x_i = 1 | \theta) = g_i + (1 - g_i)\Phi[p_i(\theta - d_i)] \qquad (6)$$

con $i = 1,2,...,\quad p_i > 0\,,\ 0 \le g_i < 1$

En la expresión (6), Xi es la puntuación de la i-ésima pregunta (1 o 0 para una respuesta correcta o incorrecta respectivamente) y la tupla compuesta por (d_i, p_i, g_i) son los parámetros que caracterizan a cada pregunta: **nivel de dificultad** (d_i), **índice de discriminación** (p_i) y **factor de adivinanza** (g_i).

La expresión (6) considerada como función de θ se denomina **curva característica de la pregunta**.

El parámetro θ, nivel de conocimiento del alumno, es desconocido y el problema se centra en la estimación de su nuevo valor, tras haber seleccionado aquella pregunta cuya tupla (d_i, p_i, g_i) es la más adecuada para el valor actual de θ.

Tanto para la elección de la pregunta a plantear al alumno, como para el cálculo de la nueva estimación del nivel de conocimiento, se utiliza un acercamiento Bayesiano, adoptando que θ sigue una distribución normal $N(M_0, V_0)$.

Intuitivamente, con la expresión (6) se intenta poner de manifiesto que en los tests adaptativos se obtiene poca información sobre un alumno si se le plantean preguntas muy fáciles o muy difíciles ya que la respuesta sería predecible, de ahí que se intenten plantear preguntas cuya dificultad sea cercana a la estimación actual del nivel de conocimiento del alumno (θ). Por tanto, parece razonable que se elija aquella pregunta con “g tan pequeña como sea posible” y con “p tan grande como sea posible”.

El modelo normal fue elegido por Owen en su demostración, por la mayor tratabilidad de los resultados obtenidos.

4.4.2.2.1. Procedimiento de actualización

El procedimiento de actualización Bayesiano se lleva a cabo con: F como el conjunto de distribuciones normales, una distribución normal a priori y una relación de cercanía definida por la "misma media y varianza".

Así, si $\theta \sim N(M_0,V_0)$ a priori y M_1, V_1 son la media y la varianza posterior a la observación del suceso X_1, se actualiza la nueva estimación del nivel de conocimiento, y se obtiene la nueva distribución a priori $N(M_1, V_1)$. Este proceso realizado recursivamente, genera una secuencia de distribuciones normales de la forma:

$$F_n = \Phi\left[(q - M_n)/V_n^{\frac{1}{2}} \right] \qquad \text{con } n = 0,1,...$$

4.4.2.2.2. Elegir el estimador

La media, M_n, se toma como la nueva estimación de θ (o actual puntuación) entre las observaciones (respuestas a las preguntas planteadas) X_n y X_{n+1}.

4.4.2.2.3. Elegir la pregunta a plantear al alumno.

La dificultad de la n-ésima pregunta se elige de modo que se satisfaga la siguiente condición:

$$|d_n - M_{n-1}| < d$$

con $n = 1,2,...$ y δ una constante suficientemente pequeña que hay que fijar.

Por tanto, se supone que para todas las preguntas se cumple:

$$0 < p' \le p_n \le p'', \qquad 0 \le k' \le g_n \big/ (1 - g_n) \le k'' \qquad \text{con } n = 1,2,...$$

y aunque no supone ninguna restricción sobre $\{(p_n, g_n)\}$ para probar la convergencia de $\{M_n\}$, parece ser que la convergencia será más rápida si se maximiza p_n y se minimiza g_n en cada paso.

El test termina cuando V_n es lo suficientemente pequeña, siendo considerada F_n como la distribución a posteriori de θ.

4.4.2.2.4. Actualización de la estimación del parámetro θ.

La distribución θ posterior a la respuesta X de una pregunta definida por la tupla (d, p, g), se obtiene a partir de la distribución a priori $\theta \sim N(M_0, V_0)$, aplicando el *teorema de Bayes*.

Si el alumno responde correctamente a la pregunta, y de acuerdo con el teorema de Bayes, la probabilidad a posteriori será:

$$P(\theta \mid 1) = \frac{1}{A}\{g + (1-g)\Phi[p(\theta - d)]\}V_0^{-\frac{1}{2}} f\left[(\theta - M_0)/V_0^{\frac{1}{2}}\right]$$

y si el alumno responde incorrectamente:

$$P(\theta \mid 0) = \left[\Phi(D)\right]^{-1}\Phi[p(d-\theta)]V_0^{-\frac{1}{2}} f\left[(\theta - M_0)/V_0^{\frac{1}{2}}\right]$$

Donde 1/A es la constante de normalización, que se obtiene de:

$$A = g + (1-g)\Phi(-D), \qquad \text{donde} \quad D = (d - M_0)/\left(p^{-2} + V_0\right)^{\frac{1}{2}}$$

Los resultados obtenidos, son las medias y varianzas a posteriori:

$$E(\theta \mid 1) = \frac{M_0 + (1-g)V_0\left(p^{-2}+V_0\right)^{-\frac{1}{2}} f(D)}{g + (1-g)\Phi(-D)}$$

$$E(\theta \mid 0) = \frac{M_0 - V_0\left(p^{-2}+V_0\right)^{-\frac{1}{2}} f(D)}{\Phi(D)}$$

$$\text{var}(\theta \mid 1) = V_0\left\{1 - (1-g)\left(1 + p^{-2}V_0^{-1}\right)^{-1} f(D)\left[\frac{(1-g)f(D)}{A} - D\right]/A\right\}$$

$$\text{var}(\theta \mid 0) = V_0\left\{1 - \left(1 + p^{-2}V_0^{-1}\right)^{-1} f(D)\left[\frac{f(D)}{\Phi(D)} + D\right]/\Phi(D)\right\}$$

Una vez calculada la nueva estimación del nivel de conocimiento, debemos elegir la siguiente pregunta a plantear al alumno, es decir, aquella que minimice la expresión:

$$l(d, p, g) = (1-A)\,\text{var}(\theta \mid 0) + A\,\text{var}(\theta \mid 1)$$

En el estudio realizado por Owen, se demuestran una serie de propiedades que ponen de manifiesto algunas de las características deseables en la realización de los tests adaptativos. Entre dichas propiedades destacan:

- $E(\theta \mid 1) > M_0, E(\theta \mid 0) < M_0$

"Las medias a posteriori, de las familias de distribuciones, disminuyen conforme el alumno responde incorrectamente a las preguntas que se le plantean y aumentan si las responde correctamente".

- $\text{var}(\theta \mid 0) < V_0$, y donde $g = 0$, $\text{var}(\theta \mid 1) < V_0$ también.

"Las varianzas a posteriori, de las familias de distribuciones, disminuyen conforme el alumno responde incorrectamente a las preguntas que se le plantean, o bien, responde correctamente a las preguntas pero sin adivinar la respuesta".

- $E(\theta \mid 0)$ se incrementa con d y si $g = 0$, también lo hace $E(\theta \mid 1)$.

"Las medias a posteriori, de las familias de distribuciones, aumentan si el alumno responde incorrectamente a las preguntas de más dificultad, o bien, si el alumno responde correctamente sin adivinar la respuesta".

- $E(\theta \mid 1)$ se decrementa con g.

"Las medias a posteriori, de las familias de distribuciones, disminuyen si se plantean al alumno preguntas con un alto factor de adivinanza".

- $l(d, p, g)$ se incrementa con g.
- $l(d, p, 0)$ se decrementa con $|d - M_0|$.

4.4.2.3. Técnicas de búsqueda eficientes.

La mayor dificultad que nos encontramos al usar las estrategias de máxima información o selección bayesiana, es la existencia de grandes bancos de preguntas. La cantidad de tiempo invertido para determinar la mejor pregunta del banco crece linealmente con el número de preguntas existentes en dicho banco. El tamaño del banco y el tiempo necesario para recorrerlo, dependerá de la velocidad de la computadora usada y de la eficiencia del procedimiento de búsqueda empleado.

Para la mayoría de las aplicaciones, un gran retardo en el planteamiento de las distintas preguntas del test es inaceptable por cuestiones de distracción, aburrimiento, etc. Para reducir este tiempo de espera, haciendo las búsquedas más eficientes, se sugieren tres procedimientos (variaciones de los mismos pueden ser usadas con cualquiera de los métodos vistos en este apartado):

1. El primer procedimiento consiste en **realizar dos búsquedas mientras que el examinando responde la pregunta actual**. Una búsqueda selecciona la pregunta que proporcionaría la máxima información sobre el nivel de conocimiento del examinando si suponemos que la pregunta que éste está contestando actualmente (al mismo tiempo que se realizan las búsquedas) es incorrecta. La otra búsqueda elegiría aquella pregunta que proporciona la máxima información suponiendo que la respuesta que el examinando da es correcta. Después de que el examinando responda a la pregunta actual, la próxima pregunta a plantear será alguna de las dos ya buscadas. Como las preguntas a plantear fueron seleccionadas mientras el examinando estaba ocupado contestando otra, la próxima pregunta debería recibirla sin retardo significativo. Este procedimiento no es recomendable si el tiempo invertido en contestar a una pregunta es muy corto ya que en lugar de disminuir el tiempo de espera, lo incrementaría.

2. El segundo procedimiento disminuye el tiempo de búsqueda **limitando el número de preguntas entre las que buscar**, por ejemplo eliminando aquellas cuya dificultad difiere en gran medida con la estimación actual del nivel de conocimiento del alumno. Aunque este procedimiento disminuye el tiempo de búsqueda, dependerá de la estimación del conocimiento realizada y del tipo de distribución de las preguntas en el banco.

3. El tercer procedimiento consiste en que **asociado al banco de preguntas exista una tabla de información**. Dicha tabla ordena las preguntas del banco en términos de la información que proporcionan a través de un gran rango de niveles de conocimiento. De este modo, el algoritmo de selección, elegirá aquella pregunta del banco que esté situada en primer lugar en la tabla de información, que aún no se haya planteado al alumno y que esté situada en la columna de la tabla referente al nivel de conocimiento más cercano a la estimación actual del conocimiento de dicho alumno. Como toda la información y cálculos se realizan antes de que se produzca la generación del test, el algoritmo de selección mientras genera el test sólo tiene que identificar una pregunta en la tabla previamente creada. Esto reduce el tiempo de presentación de la pregunta al

alumno a una constante muy pequeña, por lo que la velocidad de generación del test sería aceptable. Hay que hacer notar que la tabla de información podría ser usada para crear un test adaptativo estratificado, si fuese necesario.

4.4.3. PROCEDIMIENTOS ALTERNATIVOS PARA LA SELECCIÓN DE PREGUNTAS.

Existen variaciones de los procedimientos de selección vistos. Dos alternativas a los procedimientos clásicos que merecen ser nombradas son: el uso de tests autoadaptativos y los tests pre-diseñados (“*testlets*”).

4.4.3.1. Tests Autoadaptativos.

Es una alternativa a las técnicas de selección vistas, y que consiste en minimizar la ansiedad del alumno y maximizar su rendimiento, permitiéndole participar en la elección de la pregunta del test.

En un test autoadaptativo, el alumno recibe una pregunta del test, la contesta, el sistema le informa de si su respuesta fue o no correcta y pide al alumno cómo de difícil debe ser la siguiente pregunta que se le plantee. Para facilitar las decisiones del alumno en lo referente a la dificultad de la pregunta, las preguntas son estructuradas en grupos de dificultad o estratos como en un test adaptativo estratificado. La mayor diferencia con los tests adaptativos estratificados es que el alumno elige el estrado de dificultad, en lugar de elegirlo la computadora.

La característica principal de los tests autoadaptativos es también su mayor problema. Permitir a los alumnos diseñar un test según las dificultades de las preguntas que ellos decidan, puede hacer que los alumnos se sientan menos intimidados y su rendimiento en el test sea mayor. Al mismo tiempo, los alumnos tienen la oportunidad de elegir tests que están por debajo o por encina de su nivel de conocimiento óptimo. Un alumno brillante que no quiera responder preguntas difíciles podría responder todas las preguntas fáciles del banco. Por lo que este tipo de tests aunque reducen la ansiedad del alumno, podrían producir puntuaciones con muy bajo valor de información, en la mayoría de los casos.

4.4.3.2. Testlets.

Muchos autores han discutido los numerosos problemas que surgen en la implementación de los tests adaptativos computerizados si se usan los procedimientos clásicos de selección de preguntas ya descritos. Entre ellos, Wainer y Kiely (1987) proponen un procedimiento de selección de preguntas diseñado con el uso de segmentos de tests pre-estructurados, también llamados *testlets.*

El objetivo de los testlets es el de reducir los efectos del contexto que rodea a cada pregunta y que puede afectar a la puntuación obtenida en el test. Así, se pretende:

(a) que la cobertura del contenido del test con respecto al temario sea balanceada y lo más amplia posible.

(b) facilitar que preguntas que proporcionan información con respecto a otras, sean preguntadas en el orden adecuado.

(c) incrementar la dificultad conforme el test progresa.

Para ayudar a solucionar estos problemas, Wainer y Kiely propusieron que especialistas en la materia a evaluar creasen agrupamientos de preguntas con ciertas restricciones en el orden en que van a ser presentadas: **testlets**. Estos testlets se unen para formar un test adaptativo. Ya que el número de caminos que puede tener el test adaptativo final no es muy grande, los diseñadores de tests son capaces de determinar que no se presenten en ellos, problemas de contexto.

Por otro lado, el paradigma de los testlets reduce la exactitud del resultado del test, ya que su especificación es muy débil, puesto que se usa una técnica que ofrece al examinando menos oportunidades para recuperarse de respuestas inusuales, y limita el número total de preguntas, de un nivel de dificultad determinado, que una persona podría encontrarse. Por tanto, con esta técnica, la puntuación final resulta más informativa que con un test convencional de la misma longitud, pero es mucho menos informativa que la proporcionada por un test adaptativo que use por ejemplo, la selección de preguntas en función de la máxima información, método Bayesiano, o procedimiento estratificado.

Más importante aún que la pérdida de exactitud en el resultado proporcionado por el test con el uso de los testlets, es la pérdida de las ventajas prácticas que se obtienen del paradigma standard de los tests adaptativos. Los tests adaptativos deben

proporcionar un procedimiento de test muy eficiente y ayudar a minimizar la longitud que se necesita para el test, maximizando la eficiencia de la medida. Además, un test adaptativo proporciona un procedimiento de generación de tests seguro e inmediatamente disponible. Los tests adaptativos basados en testlets no proporcionan estas ventajas prácticas.

Con el uso de los testlets, la seguridad se convierte en un punto débil ya que aunque permite a los diseñadores de tests revisar los caminos posibles y evitar así que se produzcan conflictos de contexto en el test a generar, también hace posible que los alumnos tengan la oportunidad de intercambiar la suficiente información para hacer trampas y romper la efectividad del test.

4.4.4. USO DE RESTRICCIONES PARA MEJORAR LOS TESTS ADAPTATIVOS.

Una vez que se ha elegido el procedimiento de selección de preguntas a utilizar, hay que decidir qué tipos de restricciones son necesarias para permitir que el procedimiento de selección de preguntas sea aceptable y útil, para el propósito específico de los tests a crear. Es poco probable que un procedimiento de selección basado solamente en el método de máxima información, Bayesiano o de autoelección, generase tests que fueran aceptables para todos los grupos implicados en el desarrollo del test.

En su lugar, hay que definir un procedimiento de generación de tests adaptativos asistidos por ordenador que use restricciones. De este modo, ahora se seleccionarán aquellas preguntas que proporcionen la máxima información sobre un determinado alumno, al mismo tiempo que se satisfagan otras serie de condiciones. Las restricciones que pueden añadirse a las técnicas de selección de preguntas (aunque no están limitadas a éstas) son: selección aleatoria, balanceo de contenido, test sin repeticiones y eliminación de preguntas conflictivas.

4.4.4.1. Selección aleatoria de preguntas.

Para muchas aplicaciones de tests, dependiendo del propósito, los métodos de selección de preguntas puros pueden no proporcionar todos los requisitos que necesita la situación. Por ejemplo, los administradores de tests pueden querer asegurar la

utilización de muchos de los caminos que puedan elegirse para generar el test, disminuyendo así la posibilidad de plantear preguntas repetidas a distintos alumnos, disminuyendo la posibilidad de intercambio de información entre ellos y por tanto, la posibilidad de realizar trampas. Un procedimiento que reúne estas cualidades es la selección aleatoria de preguntas. Así, en el caso de los métodos vistos (máxima información, Bayesiano, etc.) en lugar de elegir la mejor pregunta, se podría añadir la posibilidad de elegir aleatoriamente entre las 2, 3, ..., 10 mejores preguntas resultantes de aplicar dicho mecanismo de selección. Los cálculos necesarios para llevar esta estrategia a cabo se realizan mientras el alumno está respondiendo la pregunta actual, disminuyendo así el retardo en la presentación de distintas preguntas.

Así, si se utiliza una técnica de selección aleatoria unida a cualquiera de las técnicas vistas con anterioridad, y el banco de preguntas es bastante grande, es virtualmente imposible que a dos alumnos con el mismo nivel de conocimiento se le planteen preguntas iguales en el test.

Esta seguridad permite a los tests adaptativos ser usados en situaciones donde los tests convencionales no podrían usarse. En los tests convencionales (los realizados con papel y lápiz) si una pregunta es comprometida, el test necesita ser rediseñado y reimpreso. En contraste, en los tests adaptativos, las preguntas pueden simplemente omitirse del banco de preguntas útiles.

4.4.4.2. Contenido Balanceado.

Uno de los objetivos a cumplir en el desarrollo de los tests adaptativos es el equilibrio del contenido que abarcan las preguntas del test que se genera, en función a la estructura global de la materia a evaluar. Esto llevaría a la existencia de porcentajes prefijados de preguntas de los tests procedentes, de las áreas de contenido existentes en el banco de preguntas.

Los procedimientos de selección de preguntas tradicionales dependen solamente de sus parámetros estimados para elegir la pregunta a plantear al alumno. Aunque esto da como resultado una puntuación muy informativa puede que los tests generados no sean aceptados por algunos profesores. Por ejemplo, si el test de un alumno de matemáticas contiene solamente problemas de división, éste es inadecuado para valorar el conocimiento del alumno en el campo de las matemáticas. Para mantener la

correspondencia del test con el curriculum de la materia enseñada, el procedimiento de selección debe evitar el evaluar sólo partes de este curriculum con el test. Un simple procedimiento para tener en cuenta esta restricción sería el siguiente:

La cobertura del contenido del test se especifica como porcentajes de las cuestiones del test que deberían provenir de cada una de las subáreas que constituyen el test. Así, un test de aritmética básica podría consistir en el 30% de preguntas sobre suma, 30% sobre resta, 20% sobre multiplicación y el 20% restante sobre división. Los pasos que se siguen durante la realización del test son:

1) El nivel de conocimiento provisional del alumno se calcula tras responder éste a la pregunta actual.
2) Se calcula el porcentaje de preguntas ya administradas en cada subárea del test.
3) Los porcentajes empíricos son comparados con el porcentaje prefijado por el diseñador del test en cada área. Aquella área cuya diferencia de porcentajes sea mayor es seleccionada.
4) Dentro del área seleccionada se elige una nueva pregunta a plantear al alumno, según el mecanismo de selección elegido.

Este procedimiento es realizado cada vez que una pregunta se plantea al alumno y tiene dos ventajas. La primera es que el test elige la pregunta óptima dentro de las restricciones del diseñador. Segundo, para tests adaptativos de longitud variable, el procedimiento proporciona preguntas que se acercan lo máximo posible a las restricciones dadas en las especificaciones.

La razón más importante para implementar el balanceo de contenido en los entornos educacionales, es proporcionar tests adaptativos próximos a la instrucción. Los profesores normalmente tienen ciertos objetivos y áreas de conocimiento que desean dirigir durante el curso de la instrucción. Estos objetivos a menudo difieren en dificultad y, por tanto, pueden ser evaluados de modo distinto en un test adaptativo que en un test de contenido no balanceado. Aunque podemos inferir la competencia de los alumnos en las áreas de menor dificultad a partir de su nivel de conocimiento en las áreas de mayor dificultad, es importante para los profesores poder identificar a alumnos con patrones de conocimiento peculiares, así como a aquellos que aprenden siguiendo un plan.

Un test adaptativo de contenido balanceado proporciona a los usuarios tests que representan adecuadamente cada una de las áreas de contenido incluidas. También proporciona a los diseñadores de tests un test que debería reducir el impacto de las diferencias individuales en la dimensionalidad de la respuesta.

4.4.4.3. Tests sin repeticiones.

La mayoría de las estrategias de selección vistas no evitan administrar las mismas preguntas a aquellas personas que se evalúan más de una vez. Esto significa que un estudiante que vuelve a evaluarse puede llegar a contestar preguntas repetidas más de una vez. Para asegurar que el estudiante no ha cambiado ampliamente entre las dos sesiones de test, el porcentaje de preguntas repetidas debe ser muy alto.

Para evitar "adulterar" la puntuación del estudiante debido a la repetición de preguntas, es necesario desarrollar un procedimiento de selección con restricciones, el cual elimine o minimice la repetición de preguntas. Este procedimiento debería:

(a) crear un registro de las preguntas planteadas al alumno en la sesión de test.

(b) comprobar si el alumno participó en una sesión de test anterior.

(c) seleccionar la pregunta más informativa para dicho alumno y que no le haya sido planteada aún.

Por eficiencia, éstos registros deberían contener además una fecha de expiración, para poder determinar cuando es posible el volver a plantear a un alumno preguntas que ya le han sido planteadas anteriormente. Transcurrida la fecha de expiración, estas preguntas ya planteadas podrían volver a ser planteadas al alumno.

Este procedimiento funciona muy bien en el caso de tener bancos de preguntas grandes. Si por ejemplo intentásemos este procedimiento con bancos de 100 preguntas y el test en la primera sesión fuese de 25 preguntas, el sistema plantearía preguntas con muy poca información en la segunda sesión de test, y en sucesivas sesiones nos quedaríamos sin preguntas que plantear, en el caso de no poder repetirlas.

La "adulteración" de las puntuaciones que podría resultar de la repetición de preguntas necesita ser contrastada con la pérdida de información que se puede producir por el uso de los tests sin repeticiones con bancos de preguntas no suficientemente

grandes. Por tanto, aquellos mecanismos que menos desventajas produzcan, serán los elegidos.

4.4.4.4. Eliminación de preguntas conflictivas.

Aunque la restricción de una selección sin repeticiones eliminaría el adulterar la puntuación de los alumnos entre sesiones distintas de tests, hay otro problema de "contaminación" adicional dentro de cada test generado. Las preguntas que nunca deberían ser usadas juntas en un test de papel y lápiz (preguntas conflictivas, por ejemplo: dos preguntas que proporcionan información para responder una a la otra, etc.) podrían aparecer en el orden erróneo en un test adaptativo, a menos que se añadan las restricciones adecuadas en la estrategia de selección. Un procedimiento como el de los tests sin repeticiones que guardase un registro de las preguntas planteadas en la sesión actual junto con otra serie de restricciones, resolvería este problema.

Para evitar que preguntas conflictivas aparezcan en el orden erróneo en el test adaptativo, hay que crear un registro de dichos conflictos antes de que el test adaptativo comience. Esta lista debería contener identificadores para todas las preguntas del banco que tienen conflictos con alguna otra. El procedimiento de selección elegiría las preguntas más informativas que no estuviesen en conflicto con las ya planteadas al alumno.

Como los otros procedimientos, el procedimiento para eliminar preguntas conflictivas puede afectar a la cantidad de información obtenida por el sistema y puede no ser aconsejable en caso de tener bancos de preguntas reducidos. Sin embargo, plantear preguntas que den la respuesta a otras preguntas aún no planteadas crea un problema a la hora de puntuar, difícil de cuantificar.

4.4.5. CONCLUSIONES.

Suponiendo que tenemos el software adecuado, los procedimientos de selección de máxima información y Bayesiano (junto con algún procedimiento de agilización de búsqueda) son los más adecuados hoy día. Ambos métodos optimizan la efectividad del test adaptativo y las diferencias entre ambos son mínimas. Aunque el empleo de uno u otro puede ocasionar diferencias en la elección de las preguntas, no existe evidencias que indiquen qué procedimiento proporciona el mejor conjunto de preguntas para un

determinado test. En ambos métodos, para ganar tiempo en los cálculos, la búsqueda de la nueva pregunta se hace mientras el alumno responde a la pregunta actual.

Independientemente del algoritmo elegido, éstos deben ser complementados con algunos de los algoritmos de restricciones vistos. Así, el empleo de la técnica del balanceo de contenido junto a otras, puede solucionar la necesidad del uso de los testlets sin pérdida en la exactitud del resultado o en las ventajas que ocasionan los tests adaptativos frente a los convencionales de un número preguntas prefijadas.

Como en cualquier otro procedimiento que afecte a la creación de tests adaptativos, es necesaria la existencia de bancos de 100 ó más preguntas, para obtener ventajas con respecto a los tests convencionales.

4.5. DESCRIPCIÓN DE LOS SISTEMAS EXISTENTES PARA TESTS ADAPTATIVOS

4.5.1. INTRODUCCIÓN

Algunas de las aplicaciones de los CAT hacen uso de la teoría IRT para la selección de las preguntas del test a generar y para estimar el nivel de conocimiento alcanzado por el alumno y su precisión. Estas estimaciones pueden ser usadas en conjunción con ciertas estrategias de evaluación para facilitar ciertas decisiones educacionales. Debido a la complejidad matemática del modelo IRT para estimar el valor de los parámetros asociados a cada uno de los elementos del banco de preguntas, no es práctico para los diseñadores instructores desarrollar sus propios CATs basados en este modelo. Estos problemas podrían limitar el uso de los CATs en los entornos educativos. Por esta razón, tras el modelo IRT han surgido alternativas tales como SPRT (Sequential Probability Ratio Test)y EXSPRT (Expert SPRT) que discutiremos a continuación.

4.5.2. MÉTODOS DE CLASIFICACIÓN BASADOS EN LA TEORÍA IRT

Las tres aplicaciones consideradas son:

(a) Evaluación adaptativa de la destreza, para determinar si un estudiante posee los conocimientos mínimos sobre un área determinada (AMT). Es decir, clasifica a los alumnos en dos posibles categorías: APTO y NO APTO.

(b) Evaluación adaptativa mediante asignación de grados (AGT). Clasifica a los alumnos en una serie de categorías previamente definidas. Por ejemplo: suspenso, aprobado, notable, etc.

(c) Evaluación adaptativa para estimar cambios en el nivel alcanzado previamente por los alumnos (ASRT). Determina si el conocimiento que posee un alumno varía con el tiempo y en qué medida.

4.5.2.1. AMT.

La evaluación adaptativa del conocimiento (AMT; Adaptive Mastery Testing) (Weiss & Kingsbury, 1984) determina si el nivel conocimiento estimado para un alumno está por encima o por debajo del nivel de conocimiento prefijado. Este nivel de conocimiento prefijado se determina transformando la porción del criterio de destreza especificado por el educador, a la métrica IRT, mediante el uso de la curva característica de las preguntas del banco (ICC; Item Characteristic Curve). Anteriormente al método AMT, se suponía que el nivel de conocimiento de entrada que tenía el estudiante era igual al nivel de conocimiento prefijado por el educador. Dicha suposición es bastante apropiada para AMT, donde la decisión esta basada solamente en las respuestas de los alumnos a las preguntas dadas.

Para empezar el test, se selecciona la pregunta del banco que posee el máximo nivel de información sobre el nivel prefijado. La pregunta se plantea al alumno, la respuesta del alumno se puntúa como correcta o incorrecta y el nivel alcanzado por el alumno se vuelve a estimar. El resultado es la estimación del nuevo nivel de conocimiento alcanzado por el alumno, y su intervalo de confianza (resultado de aplicar el método de máxima verosimilitud o el método bayesiano). A continuación se determina si el nivel prefijado por el educador cae dentro del intervalo de confianza del conocimiento estimado para el alumno. Si cae, se selecciona otra pregunta del banco y se opera del mismo modo. Este procedimiento continúa hasta que el intervalo de confianza no incluye el nivel prefijado, y entonces se da por **finalizado el test**. Si el límite inferior del intervalo de confianza cae por encima del nivel estimado para el

alumno, el alumno se clasifica como APTO. Por otro lado, si el límite superior del intervalo de confianza cae por debajo del nivel prefijado, el alumno es clasificado como NO APTO.

Dado un banco de tamaño finito, y un estudiante con un conocimiento de entrada cercano al nivel prefijado por el educador, la evaluación podría agotar el banco de preguntas antes de poder clasificar al alumno en APTO o NO APTO. La decisión de clasificar al alumno en esta situación, podría ser simplemente la de llevar a cabo la comprobación de si el conocimiento estimado tras la respuesta dada a la k-ésima pregunta está por debajo o por encima del nivel prefijado. Notar sin embargo, que la decisión tomada con esta regla puede no tener el mismo grado de confianza que si la hiciésemos con la anterior regla.

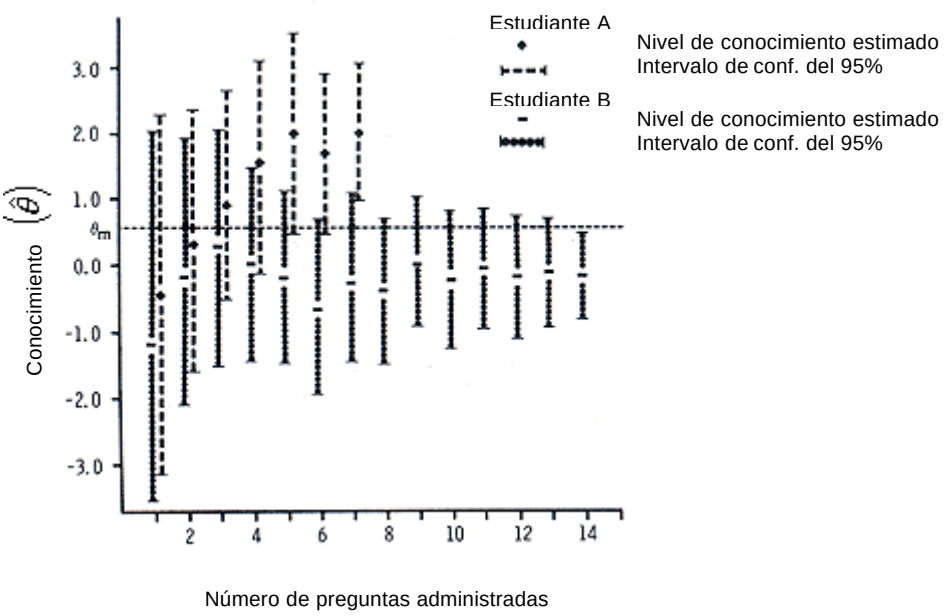


Figura.10. Ejemplo del procedimiento AMT.

La figura10 muestra el resultado de aplicar el procedimiento AMT sobre dos alumnos hipotéticos. Supongamos que asignamos un 95% de confianza a los resultados de la clasificación de los alumnos, y que el nivel de conocimiento prefijado por el educador (q_m) es de 0.50.

Para el alumno A, la estimación del conocimiento (q) después de la pregunta 1 es de 0.5 cuyo rango de confianza contiene a q_m. Por tanto, se selecciona una segunda pregunta y se le plantea al alumno. En las 6 primeras preguntas, el intervalo de

confianza contenía el nivel prefijado. Después de que el alumno respondiese a la pregunta nº 7, su nivel de conocimiento estimado y el intervalo de confianza de éste, quedaban por encima del nivel prefijado, q_m. Por tanto, el alumno A fue catalogado como APTO y el test se da por acabado.

Para el alumno B, se sigue el mismo procedimiento. Tras las primeras 13 preguntas, el intervalo de confianza de su nivel de conocimiento estimado contenía el nivel de conocimiento prefijado por el educador. La pregunta nº 14 generó un intervalo de confianza que dejó completamente por encima el nivel prefijado. En este punto, el test finaliza con el resultado de alumno NO APTO.

Observando los resultados obtenidos, la estimación para el alumno A: (2.0), está más alejada de q_m que la estimación del alumno B (-0.30). Por tanto, ha sido necesaria una estimación más precisa para el alumno B que para el alumno A. Precisión que se ha puesto de manifiesto mediante el planteamiento del doble de preguntas al estudiante B para poder llegar a una decisión.

En este método de clasificación, el punto de corte o valor umbral del nivel de conocimiento prefijado por el educador, se supone que está libre de error. Los errores de medidas están asociados en las estimaciones de los alumnos.

4.5.2.2. AGT.

Una generalización del procedimiento de longitud variable AMT es la evaluación adaptativa mediante grados (AGT; Adaptive Grading Test) (Weiss & Kingsbury, 1984), que proporciona un significado eficiente para la clasificación de los alumnos en una o más categorías. Esta generalización requiere sólo dos cambios en el procedimiento AMT:

1. Hay que establecer múltiples niveles umbrales en lugar de uno sólo como en AMT. Se necesitan N-1 niveles umbrales si la clasificación está compuesta por N grados. Como en el procedimiento dicotómico de AMT, la métrica de los niveles umbrales sobre el número-de-correctas/proporción-de-correctas es convertida a la métrica IRT mediante la curva característica del test (TCC; Test Characteristic Curve) generada para todas las preguntas previamente calibradas del banco, y con las que se generará el test adaptativo.

2. El test ha de terminar si el intervalo de confianza del conocimiento estimado del alumno cae completamente en uno de los niveles umbrales.

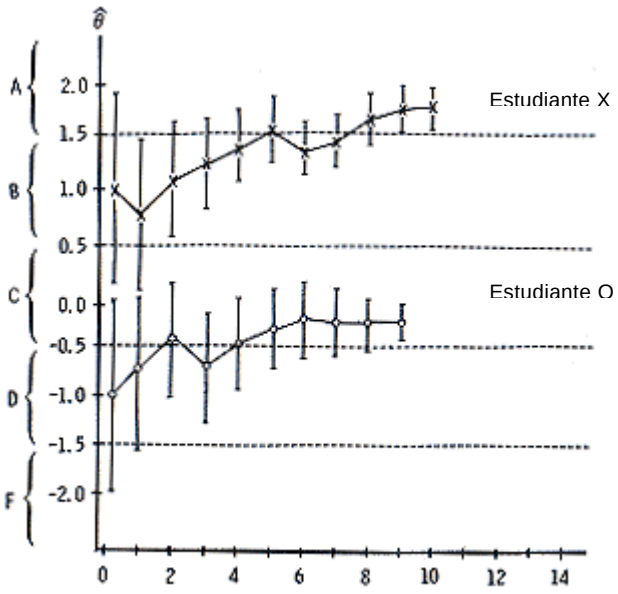


Figura 11. Ejemplo de evaluación mediante grados.

La figura 11 muestra la evaluación adaptativa mediante grados para dos estudiantes hipotéticos : X y O. Supongamos 5 grados: A, B, C, D, E y F. Previos tests para el estudiante X sugieren que se trata de un estudiante de clase "B", así que el nivel de entrada al algoritmo AGT es $q = 1$ (el punto central correspondiente al grado B). Antes de iniciar el test, hay una gran incertidumbre sobre el nivel inicial del estudiante X, por lo que el nivel de confianza dado a este estudiante va desde el punto central correspondiente al grado A (2.0), hasta cerca del punto central del grado C (aprox. 0.20).

Después de la 10^{a} pregunta, la estimación del nivel de conocimiento del alumno es de $q = 1.75$ con un intervalo de confianza [1.6, 1.9]. Como tanto q como su intervalo de confianza caen completamente dentro del grado A, el test acaba para el estudiante X con una calificación correspondiente al nivel A.

Del mismo modo se operaría con el estudiante O.

Al igual que en AMT, la longitud del test en AGT variará dependiendo de cómo de cerca está el nivel del alumno respecto de las cotas de los distintos grados. Para estudiantes cuyas estimaciones del nivel de conocimiento estén cercanas a las cotas de los grados existentes, será difícil obtener un intervalo de confianza que caiga

completamente dentro de un solo grado; y para ello, se generará un test cuya longitud estará cerca de la longitud máxima.

El tamaño de los intervalos de confianza depende de la calidad de las preguntas del banco, como reflejo de sus índices de discriminación y del patrón de respuesta del alumno. Por tanto, los alumnos que responden consistentemente de acuerdo con el modelo IRT, es decir, aquellos que responden correctamente a preguntas fáciles e incorrectamente a preguntas difíciles, tendrán intervalos de confianza pequeños y las variaciones en las estimaciones de sus niveles de conocimiento serán cada vez más pequeñas conforme el test avance. Por el contrario, estudiantes que responden inconsistentemente según el modelo IRT, es decir, aquellos que responden incorrectamente a preguntas fáciles y correctamente a preguntas difíciles, tendrán intervalos de confianza mayores y las variaciones en las sucesivas estimaciones del nivel de conocimiento no decrementarán tan rápidamente, y se necesitará mayor número de preguntas para acabar el test.

4.5.2.3. ASRT.

Además de clasificar a los alumnos según su conocimiento, los educadores también deben interesarse en saber si el conocimiento de un determinado alumno ha cambiado como efecto de la instrucción. Así, decisiones de este tipo deberían considerarse como parte esencial de la medición educacional, particularmente donde las diferencias individuales son datos importantes para el diagnóstico. Los métodos convencionales de evaluación de tests han encontrado mayores problemas para medir estos cambios individuales.

Los tests adaptativos auto-referenciados (ASRT; Adaptive Self Referencing Tests) (Weiss & Kingsbury, 1984) combinan los métodos IRT y CAT en un procedimiento coherente y flexible para la medición de los cambios producidos por la instrucción. En este procedimiento, los CAT basados en la teoría IRT son usados para obtener estimaciones del nivel alcanzado a partir de un dominio de preguntas calibradas, en ocasiones separadas por un intervalo de tiempo. El cambio es medido como la diferencia entre los niveles estimados en las dos ocasiones. Se dice que ocurre un cambio significativo cuando los intervalos de confianza basados en IRT para las dos estimaciones no se superponen.

El uso de CAT basados en IRT es ideal para la valoración del estado individual de un alumno y de su cambio. El procedimiento de los CAT selecciona y administra preguntas a cada alumno en cada ocasión de evaluación, mediante un subconjunto de preguntas que proporcionan la más precisa y eficiente medida del conocimiento. El aspecto IRT asegura que todas las estimaciones están en la misma escala sin necesidad de formas paralelas. Los datos resultantes eliminan la necesidad de comparar el nivel alcanzado por el alumno con algún nivel umbral arbitrario y proporcionan una comparación de cada alumno consigo mismo a lo largo del tiempo.

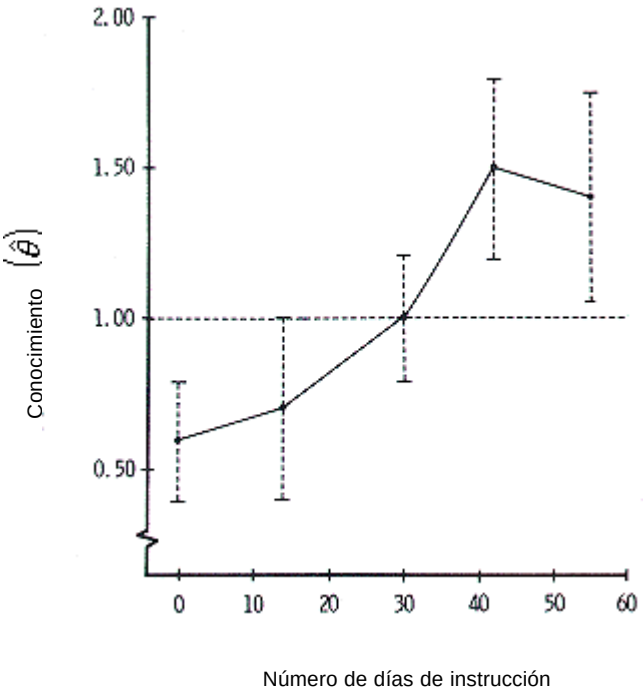


Figura 12. Ejemplo de test adaptativo auto-referenciado.

La figura 12 muestra las estimaciones del nivel de conocimiento del alumno (con los intervalos de confianza asociados) medidos en 5 ocasiones diferentes durante la instrucción: Días 0 (1er. día de la instrucción), 14, 30, 42 y 55.

En el día 0, el nivel de conocimiento estimado es de 0.60, con el 95% de confianza. La evaluación del día 14 empieza cerca del límite superior del intervalo de confianza del nivel estimado del día 0. La estimación del nivel del día 14 es sólo de 0.65 y los intervalos de confianza de los días 0 y 14 se solapan casi completamente. El día 30 la evaluación comienza cerca del 1.0 y se estima que 1.0 será el nivel que alcance ese día. Por tanto, los intervalos de confianza de los días 14 y 30 se solapan también. El día 42 el nivel estimado es de 1.50 que indica una mejora sustancial respecto de los días

anteriores. Los resultados del día 55 muestran que el nivel de conocimiento estimado el día 42 se mantiene aún.

Aunque ASRT ha sido diseñado para medir cambios en los niveles alcanzados por un individuo, puede fácilmente incorporar decisiones: APTO-NO APTO y por grados. Lo único necesario es la existencia de niveles umbrales para la clasificación que se desee hacer. Por ejemplo, la línea discontinua con $\boldsymbol{q} = 1$ en la figura 12 podría representar una clasificación APTO-NO APTO. Así, el profesor podría ver que en los días 0 y 14 el alumno era calificado de NO APTO, mientras que los días 42 y 55 el alumno es ya considerado APTO.

Ya que ASRT permite mostrar los cambios individuales en los niveles alcanzados, puede proporcionar una herramienta útil para identificar a alumnos con problemas. Por ejemplo, si un profesor usa estos tests cada semana, podría crearse un perfil del aprendizaje de cada alumno de la clase. A su vez, se podría crear el perfil de la totalidad de la clase y comparar los perfiles de cada uno de los alumnos con el perfil de la clase e identificar así que alumnos aprenden con mayor o menor rapidez, qué alumnos se estancan en ciertas partes del temario de la materia, respecto a otros que comenzaron con el mismo nivel, etc. La identificación de tales datos permite al profesor intervenir de manera adecuada cuando más se le necesita. El uso de múltiples áreas de contenido permitiría la construcción de muchos perfiles de aprendizaje para cada alumno y la inspección del patrón de estos perfiles permitiría suplir carencias específicas de cualquier materia.

4.5.3. ALTERNATIVAS PRÁCTICAS AL MODELO IRT PARA LOS CAT.

Como alternativa práctica al modelo IRT apareció el método de análisis SPRT (Welch & Frick, 1993). Sin embargo, una gran limitación de este modelo es que no toma en cuenta de manera explícita ni el índice de discriminación de los elementos del banco, ni sus grados de dificultad.

Una mejora del método SPRT fue realizada, combinando este método con el razonamiento de los sistemas expertos. El método resultante se denominó EXSPRT (Welch & Frick, 1993). EXSPRT asigna un peso a cada pregunta del banco, permitiendo así que tanto el grado de dificultad de la pregunta, como el índice de

discriminación de ésta, sean usados en la decisión del grado de conocimiento del alumno. Según el algoritmo de elección de las preguntas, EXSPRT muestra dos alternativas: EXSPRT-R (que elige las preguntas aleatoriamente) y EXSPRT-I (que elige las preguntas de manera "inteligente", es decir, la elección de la próxima pregunta está basada en su grado de utilidad. En otras palabras, la próxima pregunta es una que aún no ha sido preguntada, y con el mejor índice de discriminación existente, y que no sea incompatible con el nivel de conocimiento estimado hasta el momento).

Común a SPRT, EXSPRT-I y EXSPRT-R es la formación de porcentajes de probabilidades discretas. En lugar de suponer una medida continua como hace IRT, estos métodos clasifican a los alumnos en categorías discretas.

En EXSPRT y SPRT la probabilidad de cada categoría discreta es estimada basándose en el vector de respuesta del alumno y en la base de reglas del sistema experto. La categoría es elegida donde la probabilidad es más alta. Esto es una importante diferencia: se elige una categoría, en lugar de un punto de algo continuo (es decir, un porcentaje de corrección).

En IRT, un CAT finaliza cuando la varianza de q, y por tanto del error de la medida en el nivel de q, se hace lo bastante pequeña para satisfacer una decisión: supera o no supera el test. En EXSPRT y SPRT, un CAT finaliza cuando la probabilidad de una categoría alcanzada es suficientemente alta para satisfacer la decisión a tomar.

Las ventajas de EXSPRT y SPRT frente a IRT son las siguientes:

- Lógica más sencilla
- Muestra representativa de estudiantes más pequeña.
- Más asequible desde el punto de vista de los entornos educativos.

La Tabla 2 muestra las comparaciones de las características de los métodos citados:

Método CAT	Método de Selección de la Pregunta	Cantidad de Datos Necesitados A priori	Uso de índice de discriminación y nivel de dificultad	Fácil de implementar
SPRT	Aleatorio	Ninguno	No	Fácil
EXSPRT-R	Aleatorio	Aprox. 50	Sí	Más difícil que SPRT
EXSPRT-I	Inteligente	Aprox. 50	Sí	Más difícil que SPRT
IRT	Inteligente	Entre 200 y 1000	Sí	Muy difícil

4.5.3.1. FORMACIÓN DE LAS REGLAS DEL SISTEMA EXPERTO EN EL METODO SPRT.

Supongamos que la experiencia en evaluación de alumnos nos da información relativa a la puntuación media de los alumnos que superan o no el test. Supongamos, que los alumnos calificados como *'aptos'* obtienen en el test una puntuación de 0.85, mientras que los *'no aptos*' tienen 0.40. Desde la perspectiva de los sistemas expertos, estas condiciones son expresadas mediante reglas 'SI ...ENTONCES' (que desde el punto de vista estadístico no son sino probabilidades condicionales).

Regla 1. Si el estudiante es considerado apto entonces, la probabilidad de seleccionar una pregunta que será contestada correctamente es 0.85.

En notación matemática:

⇨ **Regla 1A.** Prob(Correcta/Apto) = 0.85 ó Prob(C/A) = 0.85

⇨ **Regla 1B.** Prob(Incorrecta/Apto) = 0.15 ó Prob(~C/A) = 0.15

Regla 2. Si el estudiante es considerado no apto, la probabilidad de seleccionar una pregunta que sea contestada correctamente es 0.40.

⇨ **Regla 2A.** Prob(C/N) = 0.40

⇨ **Regla 2B.** Prob(~C/N) = 0.60

Estas serían las 4 reglas básicas de cualquiera de las variantes del método SPRT.

En un test adaptativo que use el método SPRT, aleatoriamente se elige una pregunta del banco de preguntas y se presenta al alumno. Después de observar y evaluar la respuesta de éste, se calcula una probabilidad porcentual:

$$PR = \frac{P_{om}\, P_m^r \left(1-P_m\right)^w}{P_{om}\, P_m^r \left(1-P_m\right)^w}$$

Donde:

- PR: probabilidad porcentual.
- P_{om}: probabilidad a priori sobre considerar al alumno 'apto'.
- P_{on}: probabilidad a priori sobre considerar al alumno 'no apto'.
- P_m: probabilidad de responder correctamente siendo 'alumno apto'.
- P_n: probabilidad de responder correctamente siendo 'alumno no apto'.
- r: número de respuestas correctas.
- w: número de respuestas erróneas.

A continuación, la probabilidad porcentual calculada se compara con las siguientes **reglas de decisión**:

⇨ **Regla de decisión 1**. Si $PR \geq (1-\beta)/\alpha$, entonces elegir la hipótesis de 'alumno apto' y finalizar el test.

⇨ **Regla de decisión 2**. Si $PR \leq \beta(1-\alpha)$, entonces elegir la hipótesis de 'alumno apto' y finalizar el test.

⇨ **Regla de decisión 3**. Si $\beta/(1-\alpha) < PR < (1-\beta)/\alpha$, entonces seleccionar aleatoriamente otra pregunta.

El valor de α depende de la probabilidad inicial que asignemos de que sea falsa la decisión final de 'apto', mientras que el valor de β depende de la probabilidad que asignemos de que sea falsa la decisión final de 'no apto'. Por tanto, a valores muy pequeños de α y β, mayor longitud tendrá el test ya que se necesitarán más preguntas para poder emitir una calificación para ese intervalo de error.

En definitiva, antes de comenzar el método SPRT hay que establecer el intervalo de confianza de la decisión a la que se llegue ($1-\alpha$ y $1-\beta$). Así, si deseamos un grado de confianza del 95% deberíamos colocar $\alpha = \beta = 0.025$. También se deben fijar las probabilidades a priori de considerar a un alumno 'apto ' o 'no apto' bien sea, empíricamente o elegidas específicamente.

4.5.3.2. FORMACIÓN DE LAS REGLAS DEL SISTEMA EXPERTO EN EXSPRT.

La diferencia entre SPRT y EXSPRT es que en EXSPRT cada pregunta del banco tiene asociado un conjunto de pesos. Por tanto, ligadas a cada pregunta existen 4 reglas:

Regla i.1: Si el alumno es considerado 'apto' y se selecciona la pregunta i, entonces la probabilidad de que el alumno responda correctamente es P(Ci/A).

Regla i.2: Si el alumno es considerado 'apto' y se selecciona la pregunta i, entonces la probabilidad de que el alumno responda incorrectamente es P(~Ci/A).

Regla i.3: Si el alumno es considerado 'no apto' y se selecciona la pregunta i, entonces la probabilidad de que el alumno responda correctamente es P(Ci/N).

Regla i.4: Si el alumno es considerado 'no apto' y se selecciona la pregunta i, entonces la probabilidad de que el alumno responda incorrectamente es P(~Ci/N).

Las probabilidades de cada una de las preguntas del banco son creadas haciendo uso de datos históricos con los que se creó el banco (aprox. unos 50 alumnos). Notar que es muy importarte que un número suficiente de alumnos 'aptos' y 'no aptos' son necesarios para crear el banco (unos 25 de cada categoría).

Las fórmulas para determinar las probabilidades de repuestas correctas e incorrectas son:

$$P(Ci/A) = (\#r_{im} + 1)/(\#r_{im} + \#w_{im} + 2)$$

$$P(\sim Ci/A) = 1 - P(Ci/A)$$

$$P(Ci/N) = (\#r_{in} + 1)/(\#r_{in} + \#w_{in} + 2)$$

$$P(\sim Ci/N) = 1 - P(Ci/N)$$

Donde:

$\#r_{im}$ y $\#r_{in}$: Número de personas 'aptas' o 'no aptas' que responden a las preguntas correctamente.

$\#w_{im}$ y $\#w_{in}$: Número de personas 'aptas' o 'no aptas' que responden a las preguntas incorrectamente.

La fórmula de decisión en EXSPRT es la siguiente:

$$LR = \frac{P_{om} \prod_{i=1}^{k} P(C_i \mid M)^{s} \left[1 - P\left(C_i \middle| M\right)\right]^{f}}{P_{on} \prod_{i=1}^{k} P(C_i \mid N)^{s} \left[1 - P\left(C_i \middle| N\right)\right]^{f}}$$

Donde:

LR: Probabilidad porcentual.

P_{om}: Probabilidad a priori de que el alumno es 'apto'.

P_{on}: " " " " " 'no apto'.

Y,

$s = 1,\ f = 0$ si el elemento i es contestado correctamente.

O,

$s = 0,\ f = 1$ si el elemento i es contestado incorrectamente,

$s = 0,\ f = 0$ si el elemento i no ha sido administrado.

Las reglas de terminación del test son las mismas que en el método SPRT.

Existen dos versiones de EXSPRT: EXSPRT-R, que selecciona aleatoriamente las preguntas del test que se presentarán al alumno; y EXSPRT-I, que selecciona de manera inteligente las preguntas del test, basándose no sólo en la respuesta del alumno sino también en las propiedades de las preguntas del test que maximizan la discriminación entre APTO y NO APTO siendo éstas compatibles con el nivel estimado en ese momento para el alumno.

4.5.4. TESTS ADAPTATIVOS DE CONTENIDO EQUILIBRADO: CBAT-2.

En este apartado se pretende poner de manifiesto algunos de los resultados en la investigación, para solucionar algunas de las principales dificultades que presentan el uso de los algoritmos de tests adaptativos en los entornos de aprendizaje asistidos por ordenador. Entre esas dificultades, destacan el estudio empírico necesario para calibrar los elementos del test y la dificultad en la generación de tests de contenido equilibrado que reúnan los objetivos del profesor. Presentaremos un nuevo algoritmo, CBAT-2, (Huang, 1996) para proporcionar una solución a estos problemas.

CBAT-2 (Content Balanced Adaptive Testing v.2) es un algoritmo que solventa las dificultades comentadas.

- Genera tests que cubren las áreas definidas por el profesor (curriculum del curso).
- Elimina la necesidad de realizar un estudio empírico para calibrar las preguntas del banco.
- Selecciona las preguntas de manera inteligente (aquellas que proporciona máxima información sobre el alumno), esta forma de selección evita la formación de patrones en los tests generados, evitando así, las oportunidades de que el alumno adivine las respuestas o bien, haga trampas.
- Las preguntas pueden estar asociadas a múltiples áreas del curriculum.
- Proporciona dos niveles de valoración: valoración en cada área del curriculum, así como en el test global.

4.5.4.1. Áreas de contenido en un curriculum y en un test.

En CBAT-2, el curriculum es representado por un grafo acíclico dirigido, llamado jerarquía del currículum. La jerarquización en módulos permite reflejar el nivel de detalle desde el que se ve un concepto o entidad.

Por tanto, una jerarquía de módulos captura diferentes niveles de detalle en un tipo de red semántica. Esta red semántica identifica tanto **relaciones de agregación, como relaciones de prerrequisitos** (Collins, J.A., Greer, J.E. & Huang, 1997). La relación de agregación permite que los conceptos del más alto nivel sean divididos en

sub-componentes. Estos componentes están relacionados por *la relación lógica AND o por la relación lógica OR*, dependiendo de sí existen o no arcos que agrupen las relaciones entre ellos. De este modo, las relaciones de agregación ayudan a asegurar que los tests a generar poseen un contenido equilibrado (el definido por el diseñador en el curriculum). La relaciones de prerrequisitos ofrecen otro modo de capturar el conocimiento, ayudan a guiar el orden de generación de preguntas del test y también a disminuir la longitud de éste. También pueden capturar relaciones AND-OR.

Cada nodo del grafo, es llamado un **componente** y representa un área de contenido de cierto nivel. Cada componente de la jerarquía tiene un solo padre, excepto la raíz que representa la totalidad del curso. Un componente puede tener 1 o más componentes hijos que representan sub-áreas del curriculum.

Las **preguntas** son modeladas dentro de la jerarquía con los llamados **observadores**, es decir, entidades que poseen métodos de diagnóstico que permiten valorar el grado de corrección de la respuesta del alumno a la pregunta. Una pregunta puede estar asociada a muchos componentes de cualquier nivel del curriculum. Las preguntas asociadas con los componentes de los niveles más altos, son aquellas que requieren un conocimiento general. Las preguntas asociadas a los componentes de más bajo nivel, son aquellas que requieren un conocimiento específico sobre el curso.

Dada esta red, la evaluación puede ser realizada en cada uno de los nodos de la jerarquía, sin embargo, ya que la evaluación en un nodo puede tener efecto sobre otros nodos, **el conocimiento estimado para el nodo debe ser propagado a través de la jerarquía**. Una red Bayesiana se convierte en el método más apropiado para propagar este conocimiento. No obstante, el instructor del curso debe ser capaz de construir la red con la mínima cantidad de esfuerzo, por lo que el número de probabilidades condicionales que deba suministrar, debe ser minimizado.

4.5.4.2. Inicialización de CBAT-2.

En CBAT-2, un test valora el conocimiento del alumno en dos niveles del área de contenido del curriculum. Es decir, si se realiza un test sobre un componente del grafo que representa el curriculum del curso, sólo se preguntarán aquellas cuestiones pertenecientes a dicho componente. En cambio, si se realiza un test de todo el curso, todas las preguntas del curso pueden ser seleccionadas por el test; sin embargo, el

algoritmo es sensible a la jerarquización a nivel del curso y de los módulos pero no al nivel de conceptos.

Así, como parte de la inicialización, CBAT-2 genera un sub-curriculum para cada test específico a realizar. CBAT-2 también consulta al diseñador del test (profesor) por los pesos de todas y cada unas de las áreas del test. Por defecto, todas las áreas del test que tienen cuestiones, tienen el mismo peso.

No valoramos las áreas de contenido a todos los niveles en un test, ya que si mantenemos todos los niveles de la jerarquía del curriculum del curso, el test sería normalmente demasiado largo. En la práctica, dos niveles de jerarquización son normalmente suficientes. Si se necesita una valoración más precisa a cada nivel, entonces el estudiante debería realizar tests de cada uno de los módulos de esos niveles.

4.5.4.3. Las preguntas en CBAT-2. Parámetros.

Las preguntas del test son el elemento de más bajo nivel en la jerarquía del curriculum. Pueden tener uno o más padres, lo que refleja, que para que un alumno conteste correctamente a una pregunta, deba tener conocimientos de varias áreas del curriculum.

Las preguntas en CBAT-2 están caracterizadas por dos parámetros: **el nivel de dificultad *y* el factor de adivinanza**.

El factor de adivinanza de una pregunta se determina por la porción del número de respuestas correctas que posee una pregunta de test, y el número de respuestas posibles a dicha pregunta.

El valor del nivel de dificultad está en el rango de 0 a 1 y se obtiene combinando el valor inicial dado por el diseñador del test, con la información histórica de cada alumno.

$$diff_i = \frac{20.init_i + f_i}{20 + R_i + W_i}$$

donde: $init_i$ es el nivel de dificultad inicial dado por el profesor, la constante *20* es un factor de normalización, R_i es el nº. de veces que la pregunta i fue contestada correctamente y W_i es el nº. de veces que fue contestada erróneamente. f_i es el

acumulador de dificultad de la pregunta i y que variará cada vez que el alumno contesta a una pregunta incorrectamente.

$$f_i = \sum_{j=1}^{n} k_j . f(\boldsymbol{q}'_j)$$

Donde $n = R_j + W_i$, $\boldsymbol{q}'_j$ nivel de conocimiento del alumno cuando respondió a la pregunta la jth vez; $k_j = 0$ si la jth respuesta era correcta y $k_j = 1$ si fue incorrecta; f es una función lineal que convierte un valor $\boldsymbol{q}$ (–4 a 4) a un nivel de dificultad (0 a 1).

Conforme el test avanza, R_i y W_i crecen y el nivel de dificultad converge a $\boldsymbol{f}_i = (R_i + W_i)$. Y de este modo, puede decirse que CBAT-2 tiene cierta habilidad de aprendizaje.

4.5.4.4. Algoritmo de test.

Básicamente son 3 procedimientos: **selección de preguntas, estimación del conocimiento y puntuación**. Tales procedimientos son similares a los empleados en el algoritmo AMT (Kingsbury y Weiss –1979-) sólo que adaptados para cumplir los objetivos de CBAT-2 comentados al inicio.

4.5.4.4.1. *Selección de Preguntas.*

Abarca dos pasos: selección del componente del que procederá la pregunta y la selección de la pregunta correspondiente al componente elegido.

El componte a elegir será aleatoriamente seleccionado de entre aquellos componentes candidatos, siendo un componente candidato aquél para el que el conocimiento del alumno aún no se ha decidido. La probabilidad de elegir un componente candidato, no será la misma para todos, depende de su peso:

$$P_i = \frac{W_i}{\sum W_j / C_j es\_componente\_candidato}$$

La elección de la pregunta está basada en la cantidad de información que dicha pregunta proporcione sobre el alumno. Esta cantidad de información es calculada basada en la Curva Logística de Birnbaum donde los parámetros usados en la teoría

IRT son fijados a los valores: $a = 1.2$ y $b = g(diff_i)$; con g la función inversa de f (transforma valores de 0 a 1 a valores en el rango –4 a 4) y c el factor de adivinanza.

Una vez que la cantidad de información es calculada para cada pregunta, se selecciona aleatoriamente una, entre aquellas que proporcionen la mayor cantidad de información.

4.5.4.4.2. Estimar el nuevo conocimiento del alumno.

Una vez que el alumno responde a la pregunta planteada y el sistema determina la corrección de dicha pregunta, se actualiza el conocimiento actual del alumno, respecto al test global, y respecto al componente asociado a la pregunta. El algoritmo que emplea para el cálculo de los nuevos niveles de conocimiento del alumno y de sus intervalos de confianza, es el '*Procedimiento de Actualización Bayesiana'* de Owen (1975).

4.5.4.4.3. Puntuación.

Los procedimientos de selección de preguntas y estimación del nuevo conocimiento se repiten hasta que el test se da por finalizado. Los criterios que sigue CBAT-2 para finalizar el test son: (a) que el nivel de conocimiento sobrepase el nivel de confianza establecido por el diseñador del test, y (b) que de cada componente del test se hayan planteado al menos el mínimo número de preguntas definido también por el diseñador de test.

Una vez que el test finaliza, se usa el mismo procedimiento que en el algoritmo AMT para decidir si el alumno es o no apto. Sin embargo, el profesor debe usar con precaución esta decisión, ya que el alumno puede superar el test, no habiendo superado los niveles de confianza de alguno de los componentes del curriculum. Por tanto, si se necesita una valoración más precisa de esos componentes, el alumno debería realizar un nuevo test de ese sub-curriculum.

5. EVALUACIÓN ADAPTATIVA A TRAVÉS DE WWW

5.1. INTRODUCCIÓN

Actualmente estamos asistiendo a una nueva revolución en el mundo de la Informática y las Telecomunicaciones. Este fenómeno se llama Internet. La red de redes ha permitido la difusión de una enorme variedad de información de forma accesible a cualquier usuario. Gran parte del éxito y de la enorme expansión de Internet ha estado causada directamente por la aparición de la World Wide Web (WWW). Este servicio ha permitido la publicación de información de forma fácilmente accesible, con contenido multimedia (texto, imagen, vídeo, sonido, etc.) y fuertemente interrelacionados (las referencias en un documento no son simplemente notas sino que llevan al usuario directamente al documento referenciado).

Esta gran facilidad de uso, junto con sus capacidades multimedia hacen de la WWW un medio ideal para llevar a cabo la divulgación de material educativo y por tanto, la educación a distancia. Además, la WWW puede utilizarse para la valoración de los conocimientos impartidos. Estas dos aplicaciones de la WWW no sólo no son excluyentes sino que se complementan de forma sinérgica para formar un sistema educativo moderno y completo.

5.2. ARQUITECTURA DE UN SISTEMA BASADO EN WWW

5.2.1. EL PROTOCOLO HTTP

La base técnica de la WWW es el protocolo HTTP (HyperText Transfer Protocol). Este protocolo, desarrollado inicialmente en el CERN, se desarrolla sobre una arquitectura cliente-servidor y está diseñado para ser simple y lo más eficiente posible. Cuando un cliente desea una página determinada, conecta con el servidor que la contiene y la solicita. Si el servidor no encuentra ningún problema, la página es enviada al cliente y la conexión se termina. El protocolo, por tanto, no está basado en estados, esto es, las conexiones sucesivas entre cliente y servidor no están relacionadas de ninguna forma.

Debido a la naturaleza distribuida de la WWW se ha hecho necesaria una notación para identificar y referenciar cualquier documento que esté disponible en la WWW. Esta notación se llama URL (Uniform Resource Locator) e indica, para un documento determinado, tanto el servidor en el que está ubicado, como la ubicación concreta dentro de dicho servidor.

5.2.2. EL LENGUAJE HTML

El lenguaje HTML (HyperText Markup Language) es un subconjunto del lenguaje SGML (Structured General Markup Language). Es un lenguaje sencillo pensado para presentar información en la WWW. HTML, como su nombre indica, es un lenguaje de marcas para la creación de hipertextos. Por hipertexto entenderemos texto con una presentación agradable, con inclusión de elementos multimedia y con la presencia de hiperenlaces que permiten relacionar otras fuentes de información. Las marcas son fragmentos de texto con una estructura y sintaxis definidas que indican la estructura lógica del documento.

5.2.3. CGI (COMMON GATEWAY INTERFACE)

Unidos a la WWW destacan dos problemas fundamentales: Por un lado, el protocolo HTTP no mantiene el estado entre sucesivas conexiones del cliente, por lo que es imposible realizar interacciones que abarquen más de una página HTML. Por otra parte, las páginas HTML son documentos estáticos, esto es, su contenido es el mismo entre sucesivas peticiones, excepto en el caso de que un operador cambie el contenido de dichos documentos. Para sistemas interactivos, es necesario mostrar contenidos que no siempre están prefijados al inicio de la interacción, por lo que se hace necesario un mecanismo que permita la generación de páginas con contenido dinámico y que mantenga el estado de la interacción con el cliente a lo largo de varias páginas HTML.

Un mecanismo que solventa los dos problemas antes citados es el uso de los CGI. Estos son programas que se ejecutan en el servidor y que pueden servir para tratar información, como pasarela con una aplicación o base de datos, o para generar documentos HTML de forma dinámica.

Una de sus principales utilidades es tratar los resultados de los formularios HTML.

6. DESCRIPCIÓN DEL SISTEMA SIETTE

6.1. DESCRIPCIÓN GENERAL DE LA HERRAMIENTA.

SIETTE es un sistema de evaluación mediante tests adaptativos de contenido equilibrado, que ha sido diseñado para ser usado sobre la World Wide Web (WWW).

Usando un navegador como interfaz gráfica y, simplemente, pulsado ciertos botones se podrán crear tests del tipo "verdadero-falso", así como realizar los tests previamente creados. Además, dichos tests no sólo destacarán por ser tests adaptativos sobre WWW, sino que aceptarán como cuestiones ciertas plantillas que el usuario podrá definir y que facilitarán la creación final del test. A su vez, estas cuestiones y plantillas podrán incluir objetos multimedia, tales como: video-clips, ficheros de audio, animación y gráficos; mejorando así el aprendizaje del alumno y el modo de evaluar su conocimiento.

En definitiva, las características que ofrece el sistema SIETTE son las siguientes:

- Reduce el coste de evaluación de gran número de alumnos, la valoración es totalmente imparcial, e incremente la consistencia y precisión de los resultados obtenidos.

- Herramientas que ofrece. SIETTE suministra a los usuarios dos tipos de herramientas. Por un lado, una herramienta de edición de tests con la que se construye la base de conocimiento (especificaciones de tests, cuestiones, plantillas de cuestiones y objetos multimedia) del sistema. Por otro lado, el sistema de generación de tests, que construye el test más apropiado dependiendo del examinando.

- Confidencialidad de la información existente. El acceso tanto al editor de tests como al generador de tests, está restringido mediante palabras claves.

- Acceso concurrente a las dos herramientas que forman el sistema: el editor y el generador de tests.

- Adaptable al gusto del usuario. El formato y apariencia de las cuestiones son muy flexibles. Basta con conocer la sintaxis del lenguaje HTML para que éstas obtengan la apariencia que el creador del test considere oportuna.

- Uso de contenido multimedia. Las cuestiones del test pueden contener objetos multimedia (imágenes, vídeo, sonido, etc.) si el diseñador de test así lo desea.

- Dos tipos de tests. Tests que pueden servir tanto para que el alumno se autoevalúe como para que un profesor lleve a cabo la evaluación de sus alumnos. En un test de autoevaluación el alumno dispondrá de ayudas y explicaciones (si el profesor las facilitó al diseñar el test) a la hora de responder a cada cuestión que se le plantee.

- Accesible a través de Internet. El acceso al sistema y sus resultados están disponibles a cualquier hora, y desde cualquier lugar.

- Administración automática e "inteligente" del test. Se construyen tests adaptativos de contenido equilibrado, en los cuales el número de preguntas necesarias para que el alumno obtenga su calificación es mucho menor que en los clásicos tests de papel y lápiz, en los que todos los alumnos, independientemente de su nivel de conocimiento, tienen que responder un número fijo de cuestiones. En SIETTE, a un alumno sólo se le plantearán las preguntas adecuadas para su nivel de conocimiento, evitando así la frustración o el aburrimiento de dicho alumno. De este modo, en el sistema se han tenido en cuenta factores como: el grado de adivinanza ligado a cada pregunta del test, dificultad de las preguntas e índice de discriminación de éstas. La precisión del sistema se incrementa al aumentar el número de cuestiones que el creador del test introduce en la base de conocimiento de la que hace uso el sistema. Se favorece así, la no repetición de preguntas en la misma sesión de test, entre alumnos del mismo nivel de conocimiento, etc.

- Calificación generada por SIETTE. Como ya hemos dicho, el objetivo final del sistema SIETTE es hacer que el test haga las preguntas adecuadas para el nivel de conocimiento de un alumno y calificar a dicho alumno en APTO o NO APTO, con el menor número de preguntas posibles. Por tanto, la calificación generada por el sistema está formada por: una distribución de 10 probabilidades, una estimación del nivel de conocimiento del alumno, según esa distribución, y el intervalo de confianza asociado a dicha estimación. Intervalos de confianza pequeños, indicarán alto grado de certeza en la calificación emitida por el sistema. Por el contrario, intervalos muy grandes indicarán que el sistema posee incertidumbre a la hora de calificar al alumno.

- SIETTE crea un historial del alumno de tal modo que, si un alumno abandona un test de autoevaluación antes que de este finalice o bien lo acaba, y posteriormente

vuelve a realizar dicho test, el sistema le planteará preguntas según el nivel de conocimiento previamente estimado, sin necesidad de partir de nuevo del nivel de conocimiento inicial. En cambio, si el test es de evaluación el alumno sólo podrá realizar el test una sola vez, impidiéndose así, que pueda modificar los resultados estimados en la sesión anterior.

- Refuerzo inmediato. Los resultados del test son calculados inmediatamente y están disponibles en formato gráfico. Dichos resultados junto con otros datos sobre el alumno son almacenados para posteriores análisis sobre el perfil de aprendizaje de dicho alumno. El alumno puede ver las respuestas correctas a las preguntas que se les hace (ya que éstas no serán repetidas en la sesión actual de test) bien después de cada pregunta, o bien al finalizar el test.

6.2. DESCRIPCIÓN TÉCNICA.

6.2.1. ARQUITECTURA DEL SISTEMA SIETTE.

La *arquitectura del sistema SIETTE*, recoge los principales componentes de un test adaptativo y los agrupa en cinco módulos principales: el banco de preguntas y de especificaciones de los tests, el modelo temporal del estudiante, el editor de tests, el generador de tests, y el validador y activador de tests. La representación gráfica de la arquitectura del sistema es la que se muestra en la siguiente figura:

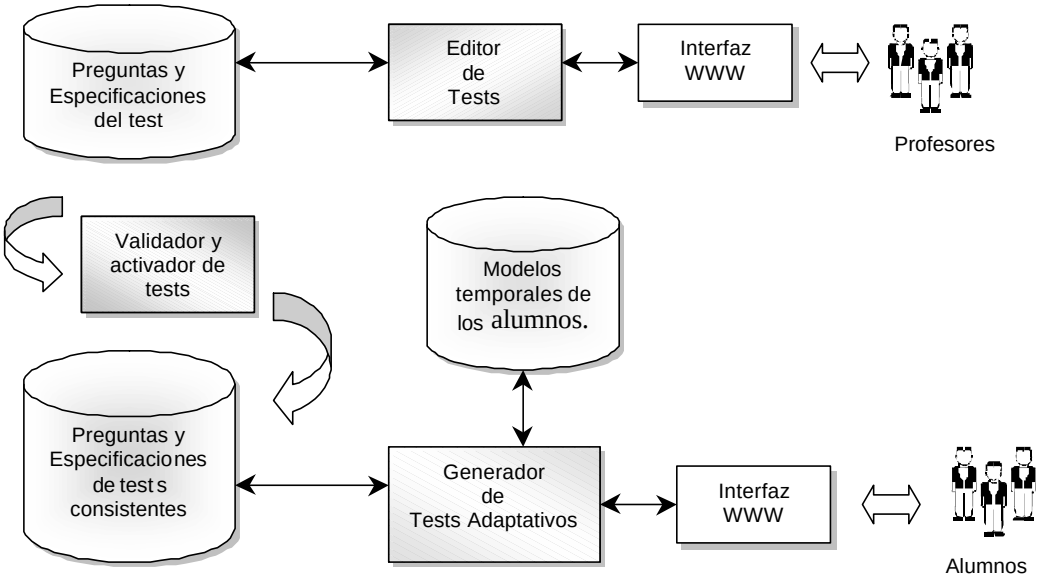


Figura 13. Arquitectura del sistema SIETTE.

El *banco de preguntas y de especificaciones de test* constituye la colección de posibles preguntas a presentar en un test, todas ellas calibradas con una serie de parámetros. Se construye haciendo uso del *módulo de edición de tests*. Dicho módulo permite a los expertos profesores, no sólo almacenar preguntas y posibles respuestas, sino que también les permite especificar el *curriculum de los tests* de las asignaturas cuya materia hay que evaluar.

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 de dicho alumno*. En el siguiente apartado veremos el algoritmo de evaluación que emplea, qué datos componen el modelo temporal del alumno y cómo se podría utilizar dicha información en cualquier STI.

Para que los tests construidos por el experto profesor sean dados a conocer a los alumnos, a través del generador de tests, sus datos deben ser validados. Sólo aquellas especificaciones de tests que cumplen los criterios mínimos de consistencia requeridos serán activadas por el módulo *validador y activador*. Dichos criterios se citarán posteriormente en este capítulo.

Tanto el módulo de edición de tests como el módulo generador de tests serán accesibles a través de la Web, constituyendo sendas interfaces de consulta y desarrollo de las distintas Bases de Conocimiento (BC) del sistema.

Por el contrario, el módulo validador y activador de tests será un proceso offline que será ejecutado 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.

6.2.2. INTERACCIÓN ENTRE LAS APLICACIONES DE USUARIO Y LAS BCS.

Tanto las aplicaciones accesibles vía Internet (editor de tests y generador de tests) como la aplicación de validación y activación de tests se encargan de interactuar con las distintas BCs, para insertar y/o modificar datos relativos tanto a los tests como a los historiales de los alumnos.

Para implementar dichas BCs se ha utilizado una base de datos (BDs) relacional denominada *Postgres* (para más información acceder a: http://www.postgresql.org) que permite su acceso vía Internet mediante una librería denominada "*Libpq"*.

Antes de continuar expondremos la arquitectura básica del sistema Postgres:

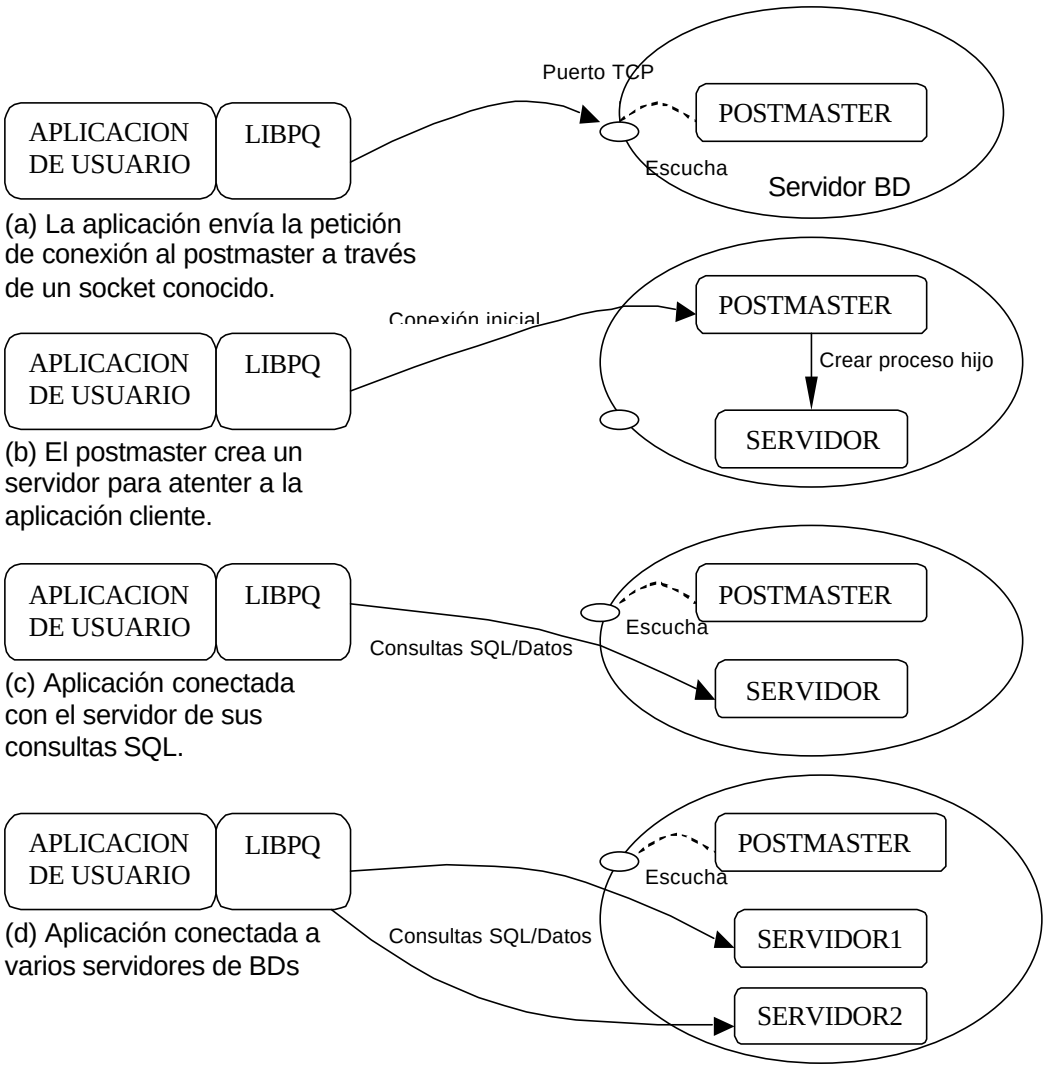


Figura 14. Arquitectura del sistema Postgres.

La BD Postgres usa un simple modelo cliente/servidor de "un proceso por usuario". La librería LIBPQ envía las peticiones de usuario sobre la red al postmaster (gráfico (a)), el cual crea un nuevo proceso servidor (gráfico (b)) y lo conecta con la aplicación cliente (gráfico (c)). A partir de este punto la aplicación y el servidor se comunican sin intervención del postmaster. Por tanto, el postmaster está siempre esperando nuevas peticiones mientras las aplicaciones prosiguen su interacción con los servidores asignados. La librería LIBPQ permite a una sola aplicación conectarse a varias BDs (varios servidores asignados - gráfico (d) -).

Una implicación de esta arquitectura es que el postmaster y los servidores creados siempre deben estar en la misma máquina (el servidor de BD), mientras que la aplicación puede ejecutarse desde cualquier lugar.

6.2.3. EDITOR DE TESTS.

El editor de tests constituye la herramienta de "elicitación" del conocimiento de los expertos profesores, almacenando dichos conocimientos en una base de datos relacional que permite su acceso a través de la WWW (motor de base de datos Postgres) mediante el uso de lenguajes de scripts, tales como el PHP/FI. Por tanto, el editor de tests está formado por diferentes scripts, todos ellos realizados en PHP/FI (http://php.iquest.net), siendo éste un lenguaje de programación embebido en el lenguaje HTML

Con la información extraída a los profesores, se crean jerarquías de tests sobre distintas asignaturas. En SIETTE, un test estará organizado de manera estructurada en *temas (tópicos) y cuestiones* que se relacionan entre sí por la existencia o no de relaciones de pertenencia definidas explícitamente por los profesores con el editor. Con la existencia de una jerarquización de la materia de la que evaluar al alumno, se pretende que el algoritmo de selección de las preguntas a plantear, tenga en cuenta esta jerarquización (Wainer & Kiely, 1987) y genere tests cuyo contenido sea equilibrado y siga la estructura definida por el diseñador en las especificaciones del test (Welch & Frick, 1993).

Por otro lado, el diseñador del test es el que calibra las preguntas del banco mediante la asignación de ciertos valores iniciales (Huang, 1996). Se elimina así la necesidad de un previo estudio empírico como sucedía en la teoría IRT (Kingsbury & Weiss, 1979).

La jerarquización del curso que se sigue en SIETTE es la siguiente:

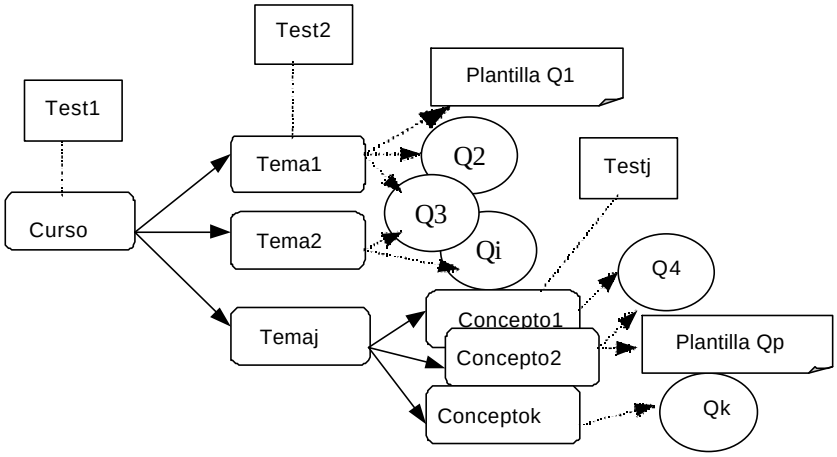


Figura 15. Vistas del sistema en diferentes tests.

Una de las grandes ventajas que ofrece este módulo de edición, es la *reutilización de los componentes*: los tests de una misma asignatura pueden compartir temas y éstos a su vez, compartir cuestiones. Además, dichos componentes pueden ser editados en HTML, por lo que ofrecen toda la flexibilidad que posee dicho lenguaje, permitiendo que los propios profesores den el formato de salida de las cuestiones que se plantearán a los alumnos durante un test.

Otra de las ventajas del curriculum de los tests, es *la presencia de contenido multimedia* tanto en los enunciados de las preguntas como en los de las posibles respuestas que acompañan a cada pregunta. Las limitaciones en la inclusión de contenido multimedia que ofrece el sistema, vienen dadas por las limitaciones que ofrece el propio lenguaje HTML como lenguaje de edición y por los navegadores actuales como intérpretes de dicho lenguaje. Con la inclusión de tales contenidos, el número de materias que se pueden evaluar con este sistema, crece sustancialmente, frente a los tests clásicos y electrónicos (la mayoría textuales y estáticos).

El mecanismo para el almacenamiento del contenido multimedia a través de la Web está basado en la *RFC 1867: "Form-based File Upload in HTML"* (Nebel & Masinter, 1995).

Uno de los puntos importantes desarrollados en SIETTE es la posibilidad de crear *plantillas de preguntas y de respuestas*. Es decir, extender el lenguaje HTML para editar los tests, de modo que en lugar de tener que introducir las preguntas una a una, se puedan definir plantillas de preguntas que se instanciarán dinámicamente si alguna de ellas es elegida por el sistema como posible pregunta a plantear al alumno. La definición de una plantilla puede dar lugar a N posibles preguntas en el test, de las que

el sistema, aleatoriamente elegirá una. Por tanto, con la definición de plantillas se minimiza aún más el riesgo de plantear preguntas repetidas, de que el alumno pueda llegar a memorizar las preguntas existentes en el banco y de que los profesores se cansen de introducir preguntas que sólo varían en el contenido pero no en los parámetros y formato de las mismas.

No se ha creado un lenguaje específico para el desarrollo de plantillas de preguntas, sino que se ha utilizado el propio lenguaje PHP/FI como lenguaje de definición.

6.2.3.1. Las preguntas en SIETTE. Parámetros.

Las preguntas del test son los elementos de más bajo nivel de la jerarquía del currículum (ver figura 15). Están caracterizadas por tres parámetros:

- **Indice de dificultad**: Fijado por el diseñador del test al crear la pregunta. Está en el rango 0-10 ya que SIETTE mide el nivel de conocimiento del alumno en 11 categorías: del Nivel 0 al Nivel 10.

- **Factor de adivinanza**: El factor de adivinanza de una pregunta se determina como la porción del número de respuestas correctas que posee una pregunta de test, y el número de respuestas posibles a dicha pregunta.

- **Indice de discriminación**: Es uno de los parámetros más difíciles y costosos de calibrar en los tests adaptativos. Basándonos en el estudio de Kingsbury y Weiss (1979) hemos inicializado dicho parámetro a la constante 1.2.

Con estos tres parámetros y aplicando un método puramente estadístico [Owen 1975], se construye para cada pregunta una curva característica constituida por 11 puntos, los 11 niveles de conocimiento.

$$P(X_i = 1 | \theta) = g_i + (1 - g_i)\Phi[p_i(\theta - d_i)]$$

donde X_i es la puntuación de la respuesta a la i-ésima pregunta (0 para incorrecta y 1 para correcta), la tupla (d_i, p_i, g_i) define los parámetros de la pregunta: nivel de dificultad, índice de discriminación y factor de adivinanza respectivamente, y Φ es la función de frecuencias acumuladas.

La expresión anterior nos da la probabilidad de responder correctamente a la pregunta i-ésima según el nivel de conocimiento estimado $\boldsymbol{q}$. Según Owen este modelo es generado asumiendo que la probabilidad de que una persona de conocimiento $\boldsymbol{q}$ sepa la respuesta correcta a una pregunta i de dificultad d_i es $P(U > d_i)$ donde $U \sim N(\boldsymbol{q}, p_i^{-2})$ y que si no sabe la respuesta correcta él puede adivinarla con una probabilidad g_i.

Aunque usamos el método Bayesiano para calibrar y seleccionar la pregunta más informativa (lo trataremos a continuación), postulamos que el conocimiento de un alumno no es infinito, sino que puede ser modelado como un conjunto finito de valores, tomando en SIETTE los valores {0..10}. Así, para calcular el nivel de conocimiento del alumno nosotros sólo analizamos, usando el método Bayesiano de Owen, parte de la distribución que encierran esos valores. Además, de este modo se facilita la computación de la nueva estimación de conocimiento y de su intervalo de confianza, después de que el alumno conteste a cada pregunta del test.

(Nota: Si la pregunta es una plantilla, todas las posibles instancias de dicha plantilla tendrán la misma curva característica ya que suponemos que al crear la plantilla lo que se quiere hacer en crear n-preguntas con el mismo nivel de dificultad y factor de adivinanza).

6.2.4. GENERADOR DE TESTS ADAPTATIVOS.

El algoritmo de generación de tests consiste en 3 procedimientos principales: (1) Selección de la pregunta a plantear según el modelo temporal del alumno. (2) Estimar el nuevo conocimiento del alumno en función de la respuesta dada por éste y con el nuevo conocimiento actualizar el modelo temporal de dicho alumno. (3) Y por último, comprobar si se cumple o no el criterio de finalización del test.

Para **seleccionar una pregunta**, se sigue el método *de selección Bayesiana* (Owen, 1975). Dicho algoritmo ha sido mejorado al añadirle las siguientes restricciones:

⇨ *Selección aleatoria*: Una de las importantes características de SIETTE es que la pregunta seleccionada, según el método Bayesiano, entre un conjunto de preguntas candidatas pueden ser una plantilla, por lo que se hará una nueva selección entre el

conjunto de instancias generadas por dicha plantilla. Esta nueva selección es aleatoria (lo mismo puede ocurrir con cada una de las respuestas posibles de la pregunta).

⇨ *Contenido balanceado*: En el sistema SIETTE, el contenido cubierto por el test es especificado por el propio diseñador del test (el profesor), como el porcentaje de cuestiones que deberían ser seleccionadas de cada una de las áreas (temas) en las que se ha estructurado dicho test. Por tanto, antes de elegir una cuestión, SIETTE debe elegir un tema candidato y para hacerlo, compara los porcentajes empíricos ya realizados por el alumno con los porcentajes preespecificados por el profesor, el tema con mayor discrepancia en esta comparación es el elegido como tema candidato. De dicho tema pasará a elegirse la cuestión apropiada.

⇨ *Tests sin repeticiones*: La estrategia de selección en SIETTE impide administrar las mismas a un alumno en la misma sesión de test. Conviene además, que dichas preguntas no le sean planteadas incluso en varias sesiones del mismo test. Esta opción, posible con SIETTE, no está disponible por el momento ya que aún no hay suficientes preguntas para ponerla en práctica. Esta última opción es recomendable cuando el banco de preguntas es suficientemente grande y proporcional al número de temas en los que están agrupadas. Se crea así, un registro de las preguntas administradas para cada alumno. Este registro tiene una fecha de expiración para indicar al sistema al cabo de cuanto tiempo dichas preguntas ya contestadas podrán volverse a plantear al alumno.

El **nuevo conocimiento del alumno y su intervalo de confianza** se estiman también por el método Bayesiano, que tiende a ser totalmente imparcial respecto el nivel de entrada prefijado para cada alumno. Con la estimación del nuevo conocimiento del alumno y con la información de las preguntas que contesta y de los temas que supera, se actualiza el *modelo temporal de dicho alumno*.

Una vez actualizado el modelo temporal del alumno se comprueba si se satisface el **criterio de finalización** del test, que consiste en cumplir las siguientes condiciones:

(1) No existir preguntas candidatas.

(2) Superar el número máximo de preguntas definido por el diseñador.

(3) Superar el mínimo nivel de conocimiento fijado por el diseñador del test, habiendo contestado un número de preguntas de cada tema mayor o igual que el establecido por el diseñador del test.

(4) La varianza a posteriori de la distribución de la estimación de conocimiento del alumno esté por debajo de un cierto umbral (fijado internamente en el sistema), habiendo contestado un número de preguntas de cada tema mayor o igual que el establecido por el diseñador del test.

Los casos (1) y (2) no deben ocurrir si el banco de preguntas es lo suficientemente grande. El caso (3) es el procedimiento abreviado de evaluación del alumno (APTO/NO APTO) haciendo caso omiso al nivel estimado. Este caso resulta recomendado si no existen suficientes preguntas o éstas están calibradas de manera errónea (es decir, el profesor asigna al azar la dificultad de las preguntas). El caso (4) es el método principal de finalización del test en los sistemas CAT cuando el banco de preguntas es suficientemente grande y está bien calibrado.

En cualquiera de los casos anteriores de finalización, el sistema devolverá no sólo la calificación sino también el intervalo de credibilidad asignado a dicha calificación. Las salidas anormales (casos (1) y (2)) tendrán intervalos de precisión grandes (mayor error en la estimación), mientras que la salida (4) será aquella con el menor intervalo y por tanto, con la estimación más precisa. La salida (3) tendrá el intervalo de confianza mínimo (asignado por el profesor al crear test) para devolver la calificación del test en el caso de un test que use un proceso abreviado de evaluación.

Mientras no se satisfagan las condiciones anteriores se siguen seleccionando preguntas, planteándolas al alumno y evaluando las respuestas dadas.

Finalmente, cuando se satisfagan alguna de las condiciones anteriores, el test acaba, los datos del modelo temporal del alumno se almacenan como conocimiento actual de éste. En el caso de ser un STI el que hace uso del sistema, el conocimiento actual del alumno pasa al módulo de diagnóstico que actualiza con él el modelo del alumno (ver capítulo "Conclusiones y Líneas futuras").

Este modulo se ha implementado como una aplicación CGI realizada en lenguaje C.

6.2.5. VALIDADOR Y ACTIVADOR DE TESTS.

La necesidad de crear este proceso dentro del sistema SIETTE ha sido debida a que la versión usada del motor de base de datos utilizado (Postgres v.6.2) no permite bloquear las tablas de la BD en modo exclusivo. Por tanto, nos encontrábamos con el

clásico problema de accesos concurrentes entre múltiples escritores (los usuarios del editor de test -los profesores-) y múltiples lectores (los usuarios del generador de test -los alumnos-) sobre la misma base de conocimiento con las preguntas y especificaciones del test.

Por dicha razón, incrementando así la seguridad de los datos existentes en la base de conocimiento del editor, decidimos que el editor interactuase con una base de conocimiento, mientras que el generador de tests interactuase con otra base de conocimiento, que embebiese a la base de conocimiento creada mediante el editor. Aprovechando la necesidad de dicho proceso, se ha hecho que ese proceso se encargue también de validar los datos suministrados de manera que la base de conocimiento con la que interactúa el generador de tests, sea una base de conocimiento consistente.

Los requisitos de consistencia que deben cumplir los datos de la base de conocimiento con las especificaciones y preguntas de los tests, para que dichos datos sean activados y de este modo pueden ser visualizados por el generador de tests, son los siguientes:

- En primer lugar se lleva a acabo la activación de las preguntas, temas y tests en la BD del editor (TEDI). Para hacerlo, primero. se deben activar las preguntas, a continuación las relaciones tests-temas y por ultimo, los test. Esto es así porque nos basamos en la activación de unos, para activar otros. Los únicos que no dependen de la activación de otro objeto son las cuestiones, por eso las activamos en primer lugar.

- Activar cuestiones: Actualiza aquellas preguntas existentes en la BD dada, activando el flag de estar o no activas para el generador de tests. Una pregunta estará activa si cumple los siguiente criterios: (1) el número de respuestas incorrectas a mostrar es mayor que 0. Esto hace que se muestre siempre una respuesta correcta y una respuesta incorrecta como mínimo, (2) el enunciado de la pregunta es un enunciado valido (distinto de vacío o de sólo blancos), (3) el número de respuestas definidas es mayor o igual que el número de respuestas que el profesor prefijó que había que mostrar. Se garantiza que para cada pregunta se mostrará al menos una respuesta correcta y una respuesta incorrecta, (4) todas las posibles respuestas a mostrar con cada pregunta tienen enunciados validos (no vacíos y distintos de blancos).

- Activar relaciones tests-temas: Actualiza aquellas relaciones entre tests y temas existentes en la BD, activando el flag de estar o no activas para el generador de tests.

Una relación test-temas estará activa si cumple los siguientes criterios: (1) cada tema tiene activas un número de preguntas mayor o igual que el mínimo numero de preguntas requerido para dicho tema.

- Activar tests: Actualiza aquellos tests existentes en la BD, activando el flag de estar o no activos para el generador de tests. Un test estará activo si cumple los siguientes criterios: (1) que todos los temas que el profesor le asoció al definir el test, estén activos, (2) requerir una puntuación mínima para superar el test mayor de 0, (3) que la fecha actual del sistema esté dentro del periodo de disponibilidad del test.

Finalmente, decir que este proceso será ejecutado por el sistema de manera periódica, que creará una copia del contenido de ambas bases de conocimiento, en forma de fichero de texto, y que cada vez que se ejecuta mantiene un fichero de sucesos, con los posibles errores producidos.

En definitiva el pseudo-código de este proceso es el siguiente:

```
/*-*-*-*-*-*-*-*-*-* PRINCIPAL -*-*-*-*-*-*-*-*-*/

int main()
{
  /* Desactiva todos los test, temas y preguntas */
  WriteToLog("Iniciando Volcado");
  WriteToLog("Realizando Copia de Seguridad de las bases de datos actuales");
  if (CopiaDeSeguridad())
  {
    exit(1);
  }
  WriteToLog("Copia de seguridad realizada");
  WriteToLog("Desactivando todos los elementos activos");
  Desactivar();
  WriteToLog("Desactivacion Terminada");

  /* Activa los elementos de test (tests, temas y preguntas) que cumplan los requisitos
de consistencia */
  /* en la base de datos del editor de tests*/
  WriteToLog("Activando todos los elementos");
  Activar();
  WriteToLog("Activacion Terminada");

  /* Sincroniza la base de datos del generador de test (SIETTE) */
  /* con la base de datos del editor (TEDI) */
  WriteToLog("Sincronizando base de datos SIETTE con base de datos TEDI");
  SincronizarBD();
  WriteToLog("Sincronizacion Terminada");
  WriteToLog("Volcado Finalizado");
  return 0;
}
/*-*-*-*-* SINCRONIZAR LAS BDs DEL EDITOR Y DEL GENERADOR -*-*-*-*-*/
void SincronizarBD()
{
   ...

   /* Conectamos con la base de datos SIETTE */
  connsiette = ConectarConBDGeneradorTest(PGHOST, PGPORT, SIETTEDATABASE);
  if (!connsiette)
  {
    WriteToLog("SincronizarBD: Error conectando con SIETTE");
    Exit_nicely(connsiette, conntedi);
  }
  /* Solicitar a la BD del generador que se quiere realizar mantenimiento del sistema y
```

```
evitar así que entren más alumnos en el sistema */
  LockDatabase(connsiette);
  /* Esperar a que los alumnos actualmente en el sistema terminen sus tests*/
  while (DatabaseInUse(connsiette))
  {
    sleep(SLEEP_TIME);
  }

  /* Abrimos la conexion con TEDI */
  conntedi = ConectarConBDEditorTest(PGHOST, PGPORT, TEDIDATABASE);
  if (!conntedi)
  {
    UnlockDatabase(connsiette);
    WriteToLog("SincronizarBD: Error conectando con TEDI");
    Exit_nicely(connsiette, conntedi);
  }
  VolcarTEDIEnSIETTE(conntedi, connsiette);
  /* Desbloquear la BD del generador de test para que de nuevo los alumnos sean
admitidos por el sistema */
  UnlockDatabase(connsiette);
  ...
}
```

7. MANUAL DE USUARIO

7.1. INSTALACIÓN Y PUESTA EN MARCHA

Para la correcta instalación del sistema y su puesta en marcha son necesarios unos requisitos previos del sistema, tanto de hardware como de software, así como seguir una serie de pasos de instalación. Vamos a revisar estos puntos con más detalle.

7.1.1. REQUISITOS EN EL SERVIDOR

7.1.1.1. Requisitos Hardware

- Ordenador IBM PC AT o compatible.
- Microprocesador Intel 80386 a 40 MHz (recomendado Pentium 166 o superior).
- 32 MB de RAM
- 50 MB de disco duro (150 MB recomendado para las bases de datos).

7.1.1.2. Requisitos Software

- Sistema Operativo Linux 2.0.0 o superior.
- Servidor WWW Apache 1.2 o superior
- Servidor de Bases de Datos SQL PostgreSQL 6.1 o superior.
- Intérprete de lenguaje PHP 2.0b12 (instalado como módulo del servidor Apache recomendado)
- Librería C estándar GNU Linux ELF 5.3.12 o superior.
- Servidor de planificación de tareas crond (opcional, pero recomendado).
- Compilador GNU C 2.7.2 (opcional)

7.1.2. REQUISITOS EN EL CLIENTE

7.1.2.1. Requisitos Software

- Navegador con soporte para HTML 2.0 o superior y JavaScript 1.0 o superior (Netscape 3.0 o superior recomendado). Además es conveniente que el sistema operativo sea basado en ventanas o soporte entornos de ventanas (Linux/XWindows, Windows 95/98, Windows 3.1, Windows NT, Mac OS, OS/2, etc.)

7.1.3. INSTALACIÓN

El sistema SIETTE viene distribuido como un único archivo en formato tar de Linux y comprimido con la utilidad GNU zip para reducir espacio. Los pasos a seguir para la instalación son:

1. **Crear un directorio para la instalación de la aplicación**. Este directorio debe estar situado en la parte del árbol de directorios accesible al servidor WWW. Copiar el archivo siette_1.0.tar.gz al directorio recién creado.
2. **Descomprimir el archivo**. Para ello teclear el comando:

   ```
   tar –zxvf siette_1.0.tar.gz
   ```

   la descompresión del archivo creará una serie de archivos y subdirectorios en el directorio de instalación.
3. **Asignar los permisos adecuados a los directorio recién creados**. Hay que asignar la propiedad de los archivos recién creados al usuario bajo el que se ejecuta el servidor WWW, así como dar todos los permisos sobre dichos archivos a este usuario.
4. **Crear un usuario PostgreSQL para el servidor WWW**. Si todavía no existe, hay que dar de alta al usuario bajo el que se ejecuta el servidor en el sistema PostgreSQL. A este usuario hay que darle permiso de creación de bases de datos.
5. **Crear las bases de datos del sistema**. Hay que crear las dos bases de datos del sistema utilizando, por ejemplo, la utilidad `createdb` de PostgreSQL. Las bases de datos deben llamarse `tedi` y `siette`.

6. **Crear las tablas del sistema**. Con el fin de hacer más cómoda la creación de las tablas del sistema se incluyen dos scripts SQL que realizan esta labor. Los scripts están situados en el subdirectorio `support` del directorio de instalación y se llaman `tedi.sql` y `siette.sql`. La ejecución de estos scripts sobre las bases de datos homónimas crea las tablas del sistema y las inicializan. Si fuera necesario habría que editar la última línea de ambos archivos para incluir el nombre de usuario bajo el que se ejecuta el servidor WWW a fin de darle los permisos necesarios sobre las tablas del sistema.

7. **Revisar la configuración del servidor WWW**. El servidor WWW debe ser configurado para la aplicación. En especial, se debe de incluir el tipo especial de script PHP en el archivo `srm.conf` y unirlo con la extensión .phtml que es la utilizada por todos los scripts PHP de la aplicación. Para ello incluir en el archivo srm.conf del servidor WWW la siguiente línea: `AddType application/x-httpd-php .phtml`
   Otro apartado a revisar de la configuración es el de los tipos MIME del servidor a fin de asociar los tipos MIME correctos con los tipos de archivo multimedia que se desean utilizar en los tests generados por SIETTE. Como mínimo habría que incluir los tipos `video/x-msvideo`, `video/quicktime`, `audio/x-wav` y `audio/basic` y asociarlos con las extensiones `.avi`, `.mov`, `.wav` y `.au`, respectivamente.

8. **Incluir en el servidor de planificación de tareas el proceso de actualización de las bases de datos del sistema**. El sistema de bases de datos necesita ser periódicamente actualizado a fin de poner a disposición de los usuarios las últimas modificaciones realizadas a los tests mediante el editor de tests. Esta tarea se realiza mediante una aplicación llamada `updatesiette`, que está situada en el subdirectorio `update` del directorio de instalación. Este proceso se puede arrancar de forma manual tecleando simplemente `updatesiette` en dicho directorio. Este proceso se debe arrancar como usuario bajo el que se ejecuta el servidor WWW. También es posible programar el servidor `crond` de forma que esta actualización se realice de forma periódica, ya sea una vez al día en horas de poco o ningún uso del sistema como en periodos más largos. Para ello se puede incluir una tarea periódica en el servidor `crond` que realice dicha actualización a los intervalos deseados.

9. **Determinar la URL de la página inicial del sistema**. Esta URL se determina conociendo el directorio raíz del servidor WWW y eliminando el path a este directorio del path de instalación. Por ejemplo, si se ha instalado la aplicación en el directorio /usr/local/etc/httpd/htdocs/siette del servidor alcor.lcc.uma.es, y el path del directorio raíz del servidor WWW es /usr/local/etc/httpd/htdocs, la URL será:

```
http://alcor.lcc.uma.es/siette/
```

7.1.4. RESOLUCIÓN DE PROBLEMAS

Problema: El sistema muestra incorrectamente los archivos con extensión .phtml.

Solución: Revisar la configuración del archivo srm.conf del servidor WWW para comprobar que se ha introducido el tipo de archivo PHP script (ver paso 7 de la instalación)

P: El sistema presenta error de acceso a la base de datos.

S: Revisar que se han creado las bases de datos y tablas correctamente (pasos 5 y 6) y que el usuario bajo el que se ejecuta el servidor WWW tiene todos los permisos para dichas bases de datos y tablas (paso 4).

P: Se han introducido tests utilizando el editor de tests y no se visualizan en SIETTE. ¿Qué puede ocurrir?

S: Comprobar que se ha actualizado el sistema empleando la utilidad `updatesiette`. Si después de empleada dicha utilidad siguen sin verse los tests, comprobar que los tests introducidos están completos y cumplen todos los requisitos para ser activados y pasar, por tanto, a ser visualizados por siette.

7.2. GUÍA DEL PROFESOR.

7.2.1. INTRODUCCIÓN

Como ya hemos comentado con anterioridad, el sistema SIETTE está formado por dos grandes módulos: el sistema de edición de la base de conocimiento del sistema SIETTE (tests, temas, cuestiones y contenido multimedia) y el sistema de generación de tests. Esta distinción refleja de algún modo, quienes serán los usuarios de cada una de

ellas. En esta sección, explicaremos cómo utilizar la herramienta de edición para crear tests que reúnan los objetivos de valoración a medir.

El editor permite a los profesores crear bancos de preguntas, de temas y de tests relativos a una asignatura; de manera cómoda y flexible. Constituye así una herramienta indispensable para la creación de la base de conocimiento del sistema de tests. Su finalidad es por una parte la de almacenar, controlar y garantizar la reusabilidad de los datos necesarios para la evaluación de cualquier alumno en una asignatura, y por otra, la de permitir al sistema de generación de tests disponer de gran número de cuestiones y de especificaciones de tests, para llevar a cabo una evaluación precisa del conocimiento del alumno.

El editor de tests es una herramienta que recoge los datos suministrados por el profesor y los almacena en el servidor. Sólo si dichos datos cumplen los requisitos mínimos para la creación de tests válidos, serán copiados de la base de conocimiento del editor a la base de conocimiento que usa el sistema de generación de tests. Se entenderá por test válido, aquél para el que existen un número mínimo de preguntas válidas para cada tema que compone el test, y que fijó el profesor. A su vez, se entenderá por preguntas válidas aquellas para las que existen una respuesta correcta y el número de respuestas incorrectas fijado por el diseñador de test (profesor). Los tests válidos que satisfagan los requisitos establecidos serán los ofrecidos por el sistema de generación de tests al resto de usuarios (preferentemente alumnos de las asignaturas).

A continuación, enumeraremos los pasos básicos para acceder al editor de test y qué opciones ofrece dicha herramienta.

Seguidamente, detallaremos las propiedades que definen cada uno de los elementos que constituyen un posible test en el sistema SIETTE (tests, temas, cuestiones y objetos multimedia – imágenes, vídeo y sonido -).

Finalmente, se mostrarán también las distintas páginas HTML a través de las cuales se solicitan los datos que definen los elementos del test.

7.2.2. PASOS BÁSICOS PARA ACCEDER AL EDITOR DE TESTS.

Para usar la herramienta de edición, el usuario debe seguir los siguientes pasos:

1) Ejecutar el navegador que tenga instalado en su PC (ver apartado “Requerimientos Software” para una explicación más detallada).

2) Conectarse a la URL donde esté instalada la aplicación (ver apartado “Instalación del sistema SIETTE”)

3) A continuación, si no hubo problemas al instalar la aplicación y no existen problemas en el funcionamiento de la red, podrá observar en su navegador la página HTML principal de la aplicación (ver figura 16). En dicha página pueden observarse dos enlaces (referencias a otras páginas HTML), cada uno de ellos irá dirigido a las dos aplicaciones con las que pueden interactuar los usuarios.

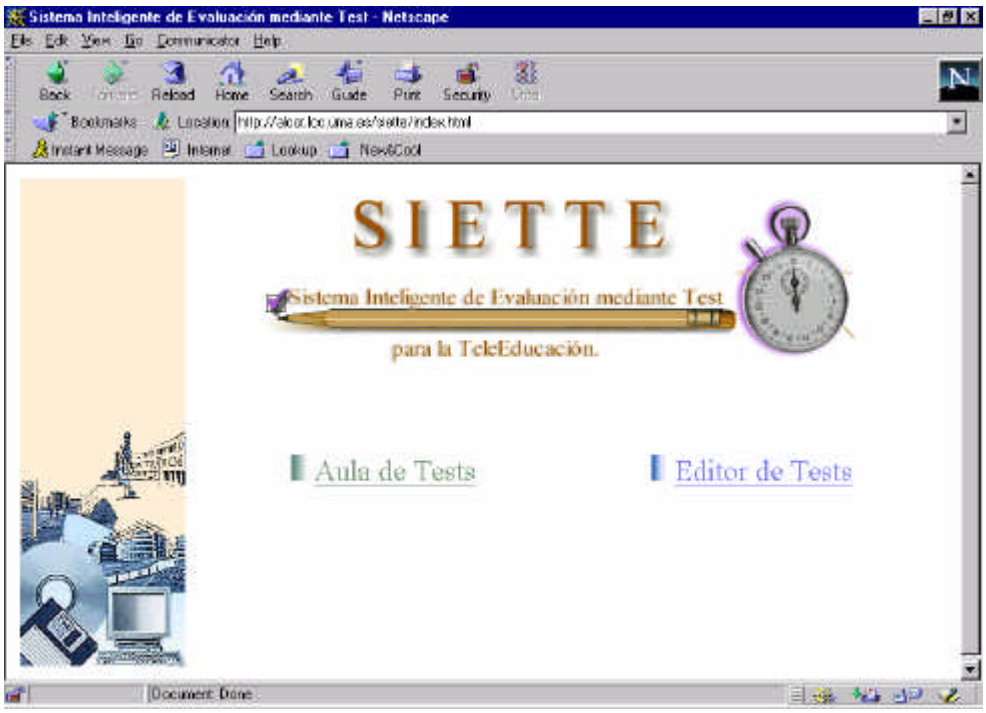


Figura 16. Pantalla principal de la aplicación.

En este caso, como lo que se quiere es crear tests, deberá hacer clic con el ratón sobre el enlace *“Editor de Tests”*.

4) Seguidamente, le aparecerá una nueva página HTML (ver figura 17) que solicitará el código de identificación de la asignatura y una clave. Ambos datos son necesarios para poder acceder al editor de tests, y le serán suministrados por el administrador de la aplicación (habitualmente el propio administrador del servidor WWW en el que está ubicada dicha aplicación.).

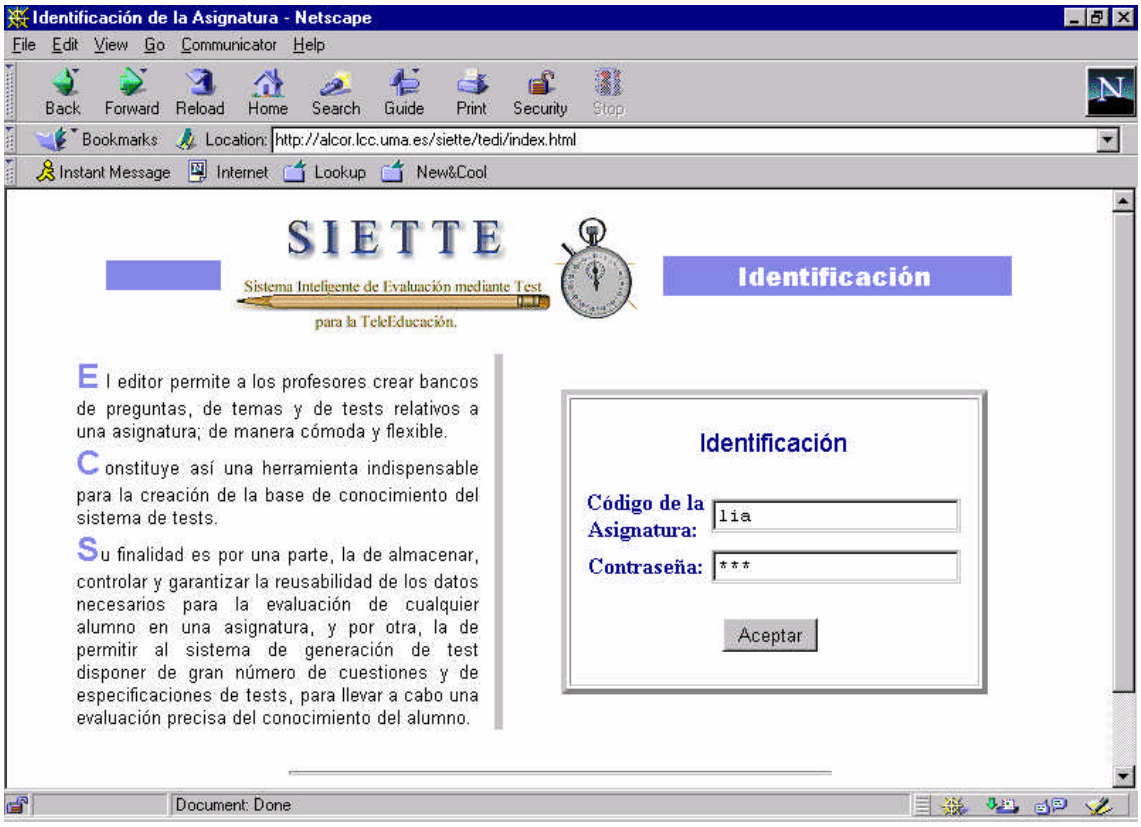


Figura 17. Página de identificación.

Esta página asegura la confidencialidad de los datos suministrados y la seguridad de los mismos.

5) Si como hemos dicho, el usuario dispone del código de identificación y de la clave de paso, el sistema mostrará el menú que ofrece la herramienta de edición, así como una breve descripción de cada una de las secciones de dicho menú (ver figura 18). En caso contrario, el sistema no permitirá el acceso al editor de tests.

7.2.3. OPCIONES QUE OFRECE EL EDITOR DE TESTS.

Las opciones que ofrece el editor de tests pueden verse en la siguiente figura:

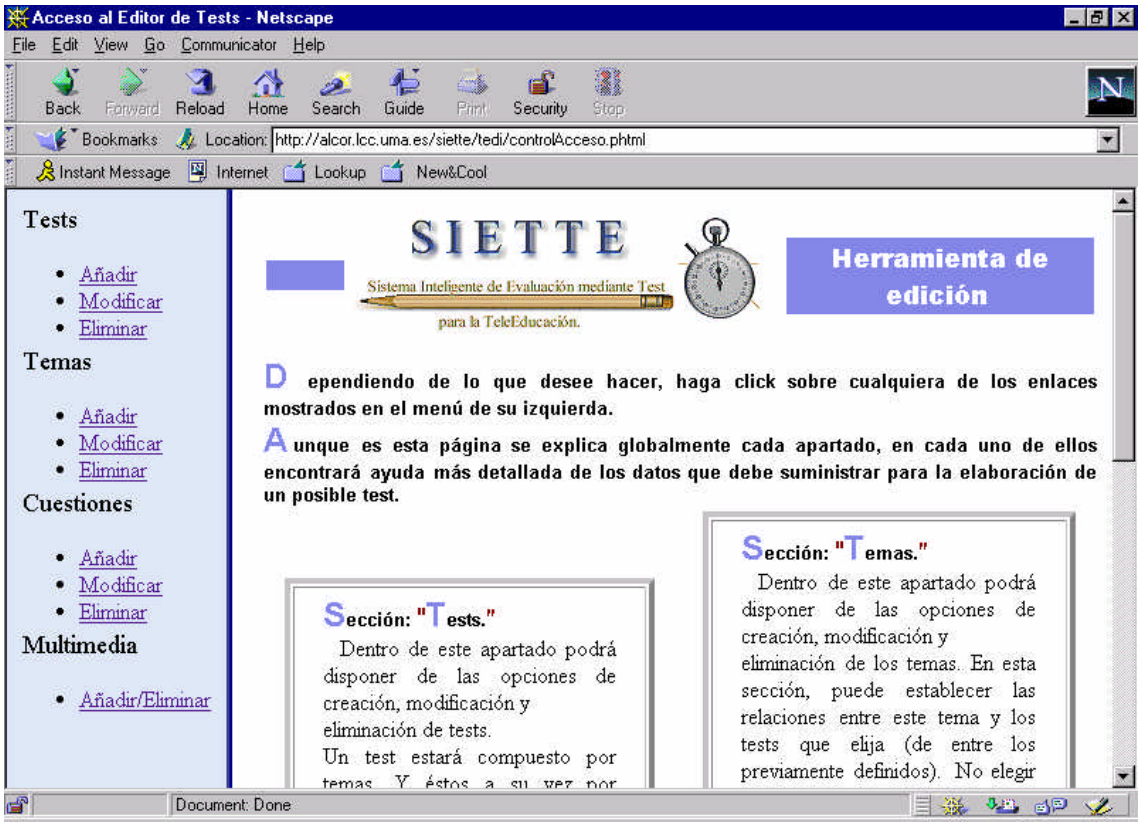


Figura 18. Menú del Editor de Tests.

- **Sección Tests**. Dentro de esta sección podrá disponer de las opciones de *creación, modificación y eliminación* de los tests pertenecientes a una asignatura. Un test estará compuesto por *temas* (unidades de agrupación de las cuestiones del test según un tópico determinado).

Cuando se añade o modifica un test, se crea una estructura en el servidor que almacena los datos suministrados para dicho test (sus propiedades generales). Cuando se elimina un test, se borran sus propiedades características y las relaciones previamente establecidas con cualquier tema perteneciente a la misma asignatura que el test. En cambio, al borrar un test no se borrarán los temas ni las cuestiones que éste contenga ya que pueden estar siendo usadas por otro test de la misma asignatura. Si se desea borrar un tema o una cuestión, el borrado deberá realizarse de manera explícita sobre dicho tema o cuestión.

- **Sección Temas**. Dentro de este apartado podrá disponer de las opciones de *creación, modificación y eliminación* de los temas. En esta sección, puede establecer las relaciones entre este tema y los tests que elija (de entre los previamente definidos). No elegir ningún test, hará que no se establezca ninguna relación y solamente almacenará los datos que usted defina para ese tema. Al eliminar un tema, se eliminarán los datos

generales de dicho tema junto con las relaciones establecidas con los tests que usan dicho tema y con las cuestiones que contiene. Para eliminar los tests o las cuestiones ha de realizarse un borrado explícito de dichos objetos.

- **Sección Cuestiones.** Dentro de este apartado podrá disponer de las opciones de *creación, modificación y eliminación* de los distintos elementos que contiene un test, es decir, de la pregunta y sus posibles respuestas. A su vez, podrá agrupar estas cuestiones en temas, si previamente los ha definido. En otro caso, simplemente se guardarán los datos introducidos y posteriormente podrá establecer las relaciones oportunas. Al eliminar una cuestión, se borrarán todos su datos (los de la pregunta y los de las posibles respuestas) y las relaciones establecidas entre dicha cuestión y cualquiera de los temas de la asignatura.

- **Sección Multimedia.** Dentro de esta sección podrá disponer de las opciones de *creación y eliminación* de los objetos multimedia (imágenes, vídeo y sonido) que pueden ser usados en las cuestiones de un determinado test. Esta opción le permitirá enviar a través de la red, y almacenar en el servidor WWW donde se ubica el sistema de generación de test, aquellos ficheros multimedia existentes en su disco local, y de los que desea hacer uso para la construcción de los posibles enunciados de las cuestiones del test.

Dependiendo de lo que desee hacer, haga clic con el ratón sobre cualquiera de los enlaces que hacen referencia a estas secciones (mostrados en el lateral izquierdo de la figura 18).

En los siguientes apartados detallaremos qué datos debe suministrar en cada sección para la elaboración de un posible test.

7.2.4. SECCIÓN TEST.

Un objeto test en el editor del sistema SIETTE no es sino el conjunto de especificaciones y propiedades generales que definirán al test, así como una serie de relaciones definidas de manera explícita entre estas propiedades y una serie de temas o tópicos de la asignatura.

Existen tres operaciones que se pueden realizar sobre un test:

(a) Creación de un nuevo test.

(b) Modificación de las especificaciones de un test existente.

(c) Eliminación de las especificaciones de uno o de varios tests.

7.2.4.1. Creación de un nuevo test.

Para crear un nuevo test, debe hacer clic sobre el enlace denominado "Añadir" de la sección "Tests" del menú del editor. Los datos que debe suministrar dentro de este apartado son los mostrados en las figuras 19 y 20:

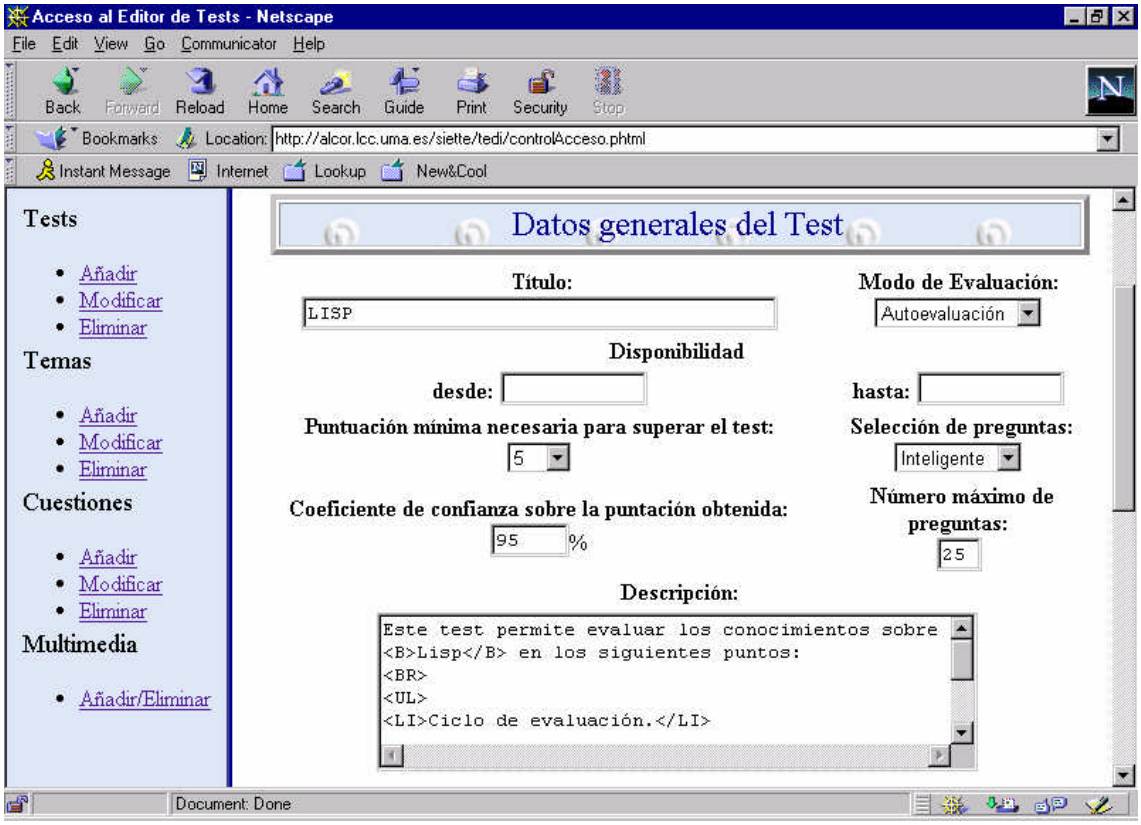


Figura 19. Datos de un test.

Los elementos que caracterizan a un test son los siguientes:

➢ **Título del test**. Nombre que recibirá el test para su identificación. Es un dato obligatorio para poder crear el test. Procure dar un nombre explícito ya que dicho título no sólo servirá para que usted identifique dicho test entre aquellos pertenecientes a la misma asignatura sino que también será el que muestre el generador de los tests a los alumnos.

➢ **Disponibilidad del test**. Periodo de tiempo durante el cual, el test podrá ser presentado a los alumnos. Transcurrido ese tiempo, el test dejará de presentarse a los alumnos pero sus datos permanecerán almacenados en el servidor. De este modo, el

profesor puede modificar o impedir que el generador presente el test a los alumnos cuando desee, sin perder la información previamente suministrada.

Si no se suministra ninguna fecha, el periodo de tiempo durante el cual el generador de tests podrá presentar dicho test a los alumnos, es ilimitado.

Si sólo se suministra la fecha relativa al intervalo inferior del periodo de disponibilidad, el test podrá ser presentado a los alumnos a partir de dicha fecha.

Si sólo se suministra la fecha relativa al intervalo final del periodo de disponibilidad, el test podrá ser presentado a los alumnos a partir de la fecha en que se cumplan los requisitos para que el test esté activo y pueda ser presentado a los alumnos (se dirán posteriormente), y permanecerá activo hasta la fecha dada (solo sí durante ese periodo se siguen cumpliendo los requisitos de consistencia para que el test esté activo).

- **Modo de evaluación**. El sistema ofrece dos modos de evaluar al alumno: **evaluación** o **autoevaluación**. La diferencia entre uno y otro es que si un test es etiquetado como evaluación, se restringirá el acceso al test a los alumnos a una sola sesión. En caso contrario, el alumno podrá realizar el test las veces que considere oportunas siempre y cuando el test esté disponible. En este segundo caso, cada vez que el alumno inicia una sesión de test, el sistema parte de la última estimación de conocimiento, para evaluar de nuevo al alumno. Se mantiene así, un modelo del alumno entre diferentes sesiones de un mismo test, evitando al alumno responder a preguntas cuyo nivel de dificultad es inferior al de la última estimación de conocimiento que hizo el sistema.

- **Selección de preguntas**. El sistema SIETTE ofrece dos modos de selección de preguntas: *aleatorio* e *inteligente*. El modo de selección **aleatorio** elige de manera aleatoria la siguiente pregunta a plantear al alumno, de entre las pertenecientes al tema seleccionado por el sistema. Este modo de selección resulta de utilidad cuando todas las cuestiones están calibradas de la misma forma y existen muchas cuestiones en la base de conocimiento del sistema. Esto es así, ya que la búsqueda resulta más eficiente al no realizar comparaciones entre las preguntas.

En el caso de selección **inteligente**, el proceso para elegir la siguiente pregunta se basa en el método Bayesiano. Es decir, se intenta elegir aquella pregunta cuyo nivel de dificultad se acerque más a la estimación que el sistema lleva hasta ese momento sobre el conocimiento del alumno. Se elegirá por tanto, aquella pregunta que minimiza la

varianza a posteriori esperada para la estimación del conocimiento del alumno. En el caso de existir varias, se elegirá una entre ellas de manera aleatoria.

- **Puntuación mínima necesaria para superar el test.** Nivel de conocimiento umbral para superar el test.

- **Coeficiente de confianza sobre la puntuación obtenida.** Grado de credibilidad asociado a la estimación de conocimiento que realiza el sistema. Mientras mayor sea el grado definido, mayor certeza tendrá la estimación calculada por el sistema. Por ejemplo, un coeficiente del 95% indica que sólo se permitirá un error del 0.05 sobre la estimación realizada. Resulta por tanto un punto clave dentro del criterio de finalización del test que lleva a cabo el generador de tests.

- **Número máximo de preguntas.** Este campo sirve para acotar el máximo número de preguntas que el sistema debe plantear a un alumno. Es necesario puesto que un alumno que intente hacer trampas, podría contestar combinaciones de preguntas de manera correcta y combinaciones de preguntas de manera errónea, de modo que el sistema no pudiese decidir con el grado de confianza asignado, cual es la calificación del alumno y por consiguiente, no dejaría de hacer preguntas. De este modo, el sistema daría a conocer al alumno, gran parte del banco de preguntas.

  Teniendo en cuenta esta consideración, el número aquí indicado debe ser un número alto y proporcional al número de preguntas existentes para ese test en la base de conocimiento. Esta opción contribuye también en el criterio de finalización del test.

- **Descripción.** Breve comentario sobre el test que se construye. Este comentario será la descripción que el sistema generador de tests presentará al alumno cuando éste desee más información sobre cierto test. Por tanto, aunque no es necesario suministrarla, su aportación es conveniente para comunicar a los alumnos cualquier aclaración, explicar los objetivos del test, etc.

- **Temas relacionados con el test.** Si existen temas ya creados en esta asignatura, aquí aparecerán sus datos. En caso contrario, sólo aparecerá un mensaje advirtiendo que si desea asociar el nuevo test a algún tema, deberá primero crearlo y luego establecer la relación. Si existen temas ya definidos, podrá asociar al test aquellos que desee, y asignar un peso a cada tema para este test. Ver figura 20.

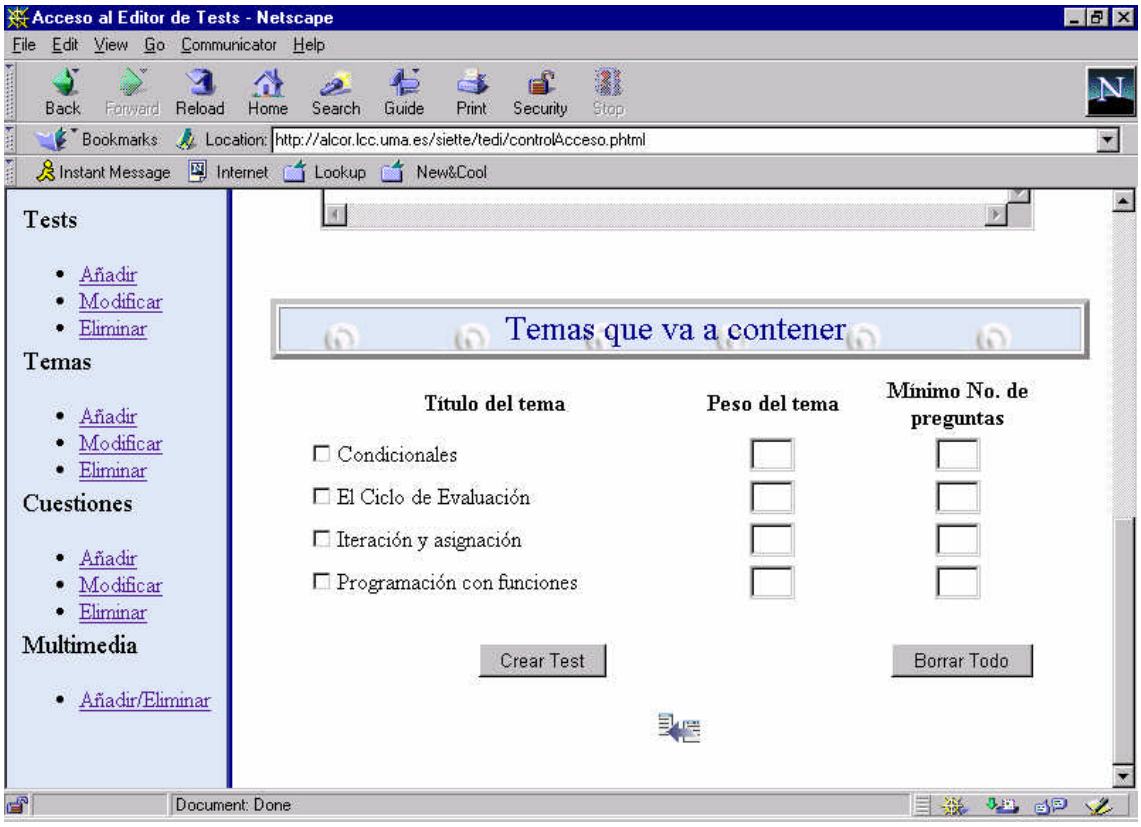


Figura 20. Relaciones que se pueden crear entre un test y varios temas.

Cuando se hayan rellenado los campos deseados del test, pulse el botón ***Crear Test***. Tanto si los datos son o no correctos, el sistema le dará un mensaje, para así proseguir o corregir los datos.

7.2.4.2. Modificación de las especificaciones de un test existente.

Si lo que se quiere es modificar alguna de las propiedades vistas en el punto anterior, hay que seleccionar la opción “Modificar” de la sección “Test”. Seleccionada esta opción, se mostrará un nuevo formulario HTML. Con los datos suministrados en dicho formulario, se realizará una búsqueda sobre todos los tests de la asignatura, que cumplan los requisitos dados por el usuario en este formulario. No aportar ningún dato en este formulario y pulsar el botón **Buscar** directamente, devolverá todos los tests existentes para una asignatura. Ver figuras 21,22 y 23.

Figura 21. Formulario HTML para la búsqueda de tests. Parte 1.

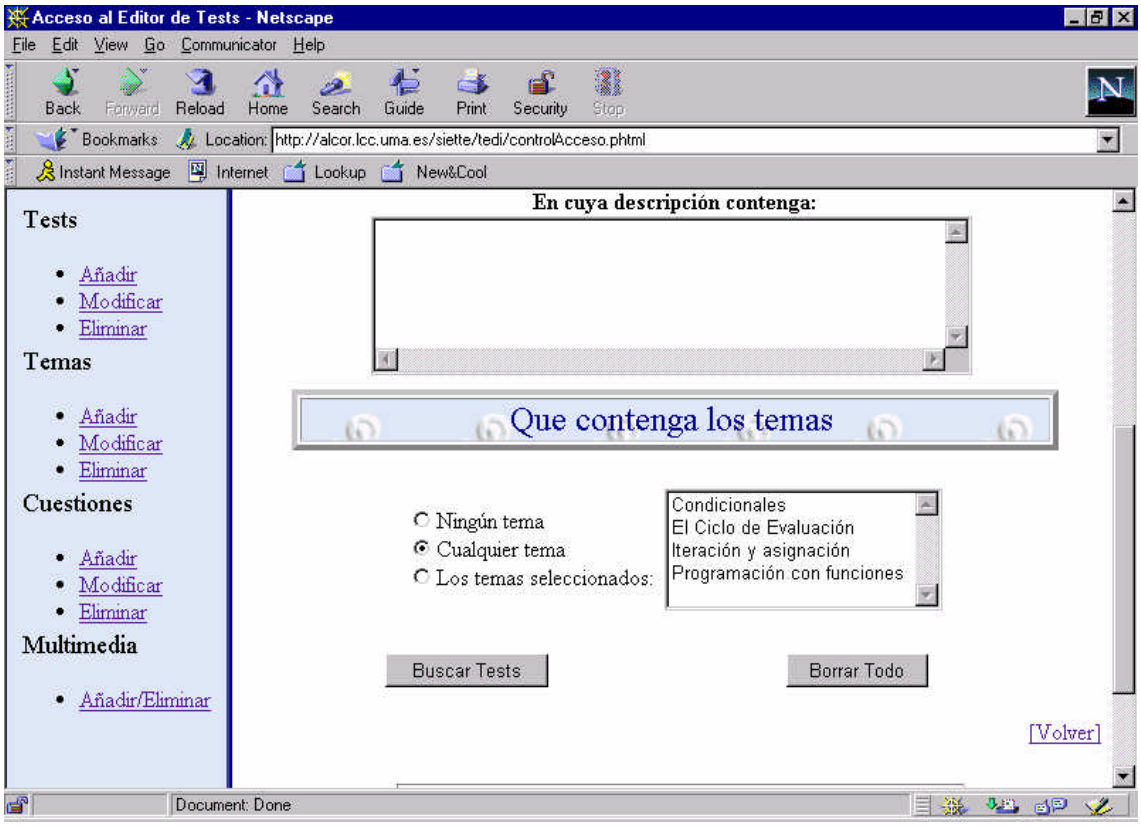


Figura 22. Formulario HTML para la búsqueda de tests. Parte 2.

Figura 23. Resultado de la búsqueda.

A partir de la lista de tests devueltos como resultado de la búsqueda, el usuario puede seleccionar aquél que desea modificar. Tras seleccionar un test, aparecerá de nuevo el formulario HTML de las figuras 19 y 20, con la salvedad de que en esta ocasión en lugar de aparecer el botón etiquetado como "Crear Test" aparecerá un botón etiquetado "Modificar Test". En este formulario aparecerán con información, aquellos campos que previamente fueron suministrados al crear el test. Basta cambiar cualquier dato y pulsar el botón ***Modificar Test*** para que los cambios sean aplicados y guardados.

7.2.4.3. Eliminación de las especificaciones de uno o de varios tests.

Del mismo modo que en el punto anterior, si el usuario selecciona del menú del editor la opción “Eliminar” de la sección “Tests”, aparecerá en primer lugar el formulario de búsqueda mostrado en las figuras 21 y 22. Como resultado de la búsqueda se devuelve una lista con aquellos tests que cumplen los requisitos fijados en dicho formulario. Ver figura 24.

Figura 24. Resultado de la búsqueda para eliminar un test.

Como puede verse en la figura 24, ahora el usuario puede:

- Borrar el test que se desee (antes de realizar el borrado el sistema pedirá confirmación al usuario).
- Borrar todos los tests mostrados (también se pedirá confirmación).
- Hacer clic sobre cualquiera de los enlaces y ver información sobre ese test. Dicha información será de sólo lectura. Tras ver la información que se suministró con anterioridad para ese test, el usuario podrá borrarlo si así lo desea pulsando el botón ***Eliminar Test***.

7.2.5. SECCIÓN TEMAS.

Un tema en el editor del sistema SIETTE es una unidad de agrupación de las cuestiones que formarán el test según un tópico determinado.

Al igual que en la sección tests, el sistema ofrece tres operaciones que se pueden realizar sobre un tema:

(d) Creación de un nuevo tema.

(e) Modificación de las especificaciones de un tema ya existente.

(f) Eliminación de las especificaciones de uno o de varios temas.

7.2.5.1. Creación de un nuevo tema.

Las propiedades que caracterizan un tema son las que se muestran en las figuras 25 y 26.

Figura 25. Datos de un tema. Datos generales y tests con los que está relacionado.

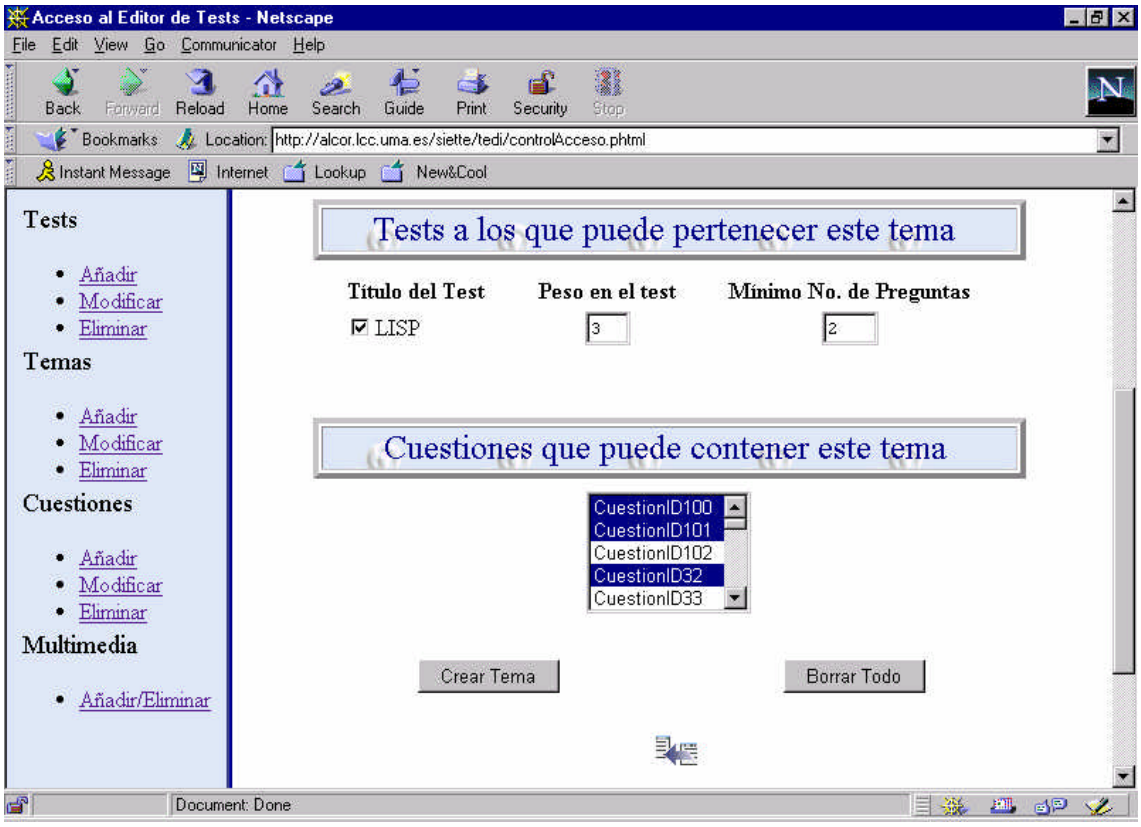


Figura 26. Datos de un tema. Cuestiones relacionadas con el tema.

Es decir, un tema está definido por un título (obligatorio para crear el tema) y un conjunto de relaciones con las cuestiones de la asignatura que pertenecerán a dicho tema. A su vez, al crear el tema, se puede asignar éste a los tests previamente definidos y asignar según el test, el peso y el número mínimo de preguntas que el sistema de generación de tests debe de plantear a los alumnos.

El peso de un tema en un test indicará la importancia de ese tema en el test con el que se asocia. Por tanto, el sistema generador de tests planteará más preguntas de los temas de más peso aunque hará siempre el mínimo número de preguntas asociado a cada tema antes de finalizar el test. De este modo, se garantizará que el contenido del test sea equilibrado según las preferencias de cada profesor.

Para definir la relación del tema con las cuestiones que va a contener, basta con seleccionar aquellas cuestiones ya definidas que aparecen en la lista de la figura 26. Para seleccionar más de una cuestión debe mantener pulsada la tecla *Ctrl* mientras hace clic sobre los nombres de las cuestiones con el ratón. Si desea seleccionar un grupo de cuestiones seguidas en la lista, debe hacer clic sobre la primera cuestión de la lista, situarse en la última cuestión del grupo a seleccionar y teniendo pulsada la tecla *shift*,

hacer clic sobre la última cuestión del grupo. Para deseleccionar alguna cuestión, debe mantener pulsada la tecla *Ctrl* mientras hace clic sobre la cuestión deseada.

Después de aportar los datos solicitados, basta con pulsar el botón ***Crear Tema*** para que el sistema almacene los datos dados. En caso de error, el sistema mostrará un mensaje para que el usuario subsane dicho error.

7.2.5.2. Modificación de los datos de un tema.

Para modificar un tema, hay que decir al sistema cuál es el tema a modificar. Para hacer esto, se tiene que:

1) Hacer clic sobre la referencia "Modificar" de la sección "Temas" del menú del editor. Aparecerá un formulario de búsqueda en el que se deben especificar los datos del tema o temas a modificar. Ver figura 27.

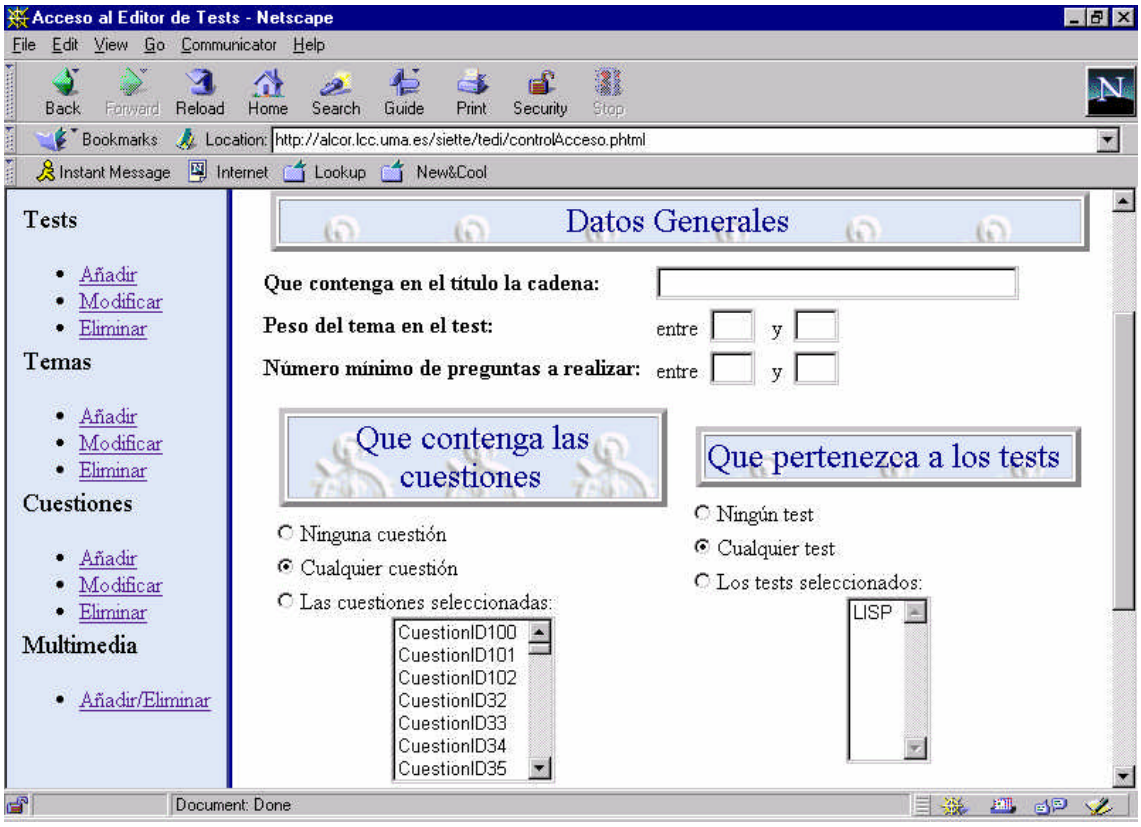


Figura 27. Formulario para realizar búsqueda entre los temas de una asignatura.

2) Realizar la búsqueda de los temas a modificar. Al final del formulario aparece un botón **Buscar.** Tras pulsarlo, se mostrará los temas encontrados en la asignatura y que cumplan las restricciones dadas a través del formulario de la figura 27. Ver figura 28.

Figura 28. Resultado de la búsqueda de los temas de la asignatura “Lisp”.

3) Seleccionando cualquiera de los temas de la lista de la figura 28, se mostrará el formulario de las figuras 25 y 26, cuyos campos estarán ahora rellenos con la información suministrada anteriormente y en lugar de aparecer el botón *Crear Tema*, aparecerá el botón ***Modificar Tema***. Pulsando este botón cualquier dato que hubiese sido cambiado, se hará efectivo.

7.2.5.3. Eliminación de los datos de un tema.

De nuevo, si se elige esta opción, se hará uso del formulario de búsqueda para poder seleccionar el tema a eliminar de la base de conocimiento del editor de tests. A diferencia de la opción *Modificar Tema* mostrada en el punto anterior, en este caso el resultado de la búsqueda permite:

1) Borrar el tema o los temas que se hayan encontrado. Se produce tras pulsar los botones ***Borrar*** o ***Borrar Todos***. En ambos casos, se pedirá confirmación al usuario antes de realizar el borrado de los datos. Ver figura 29.

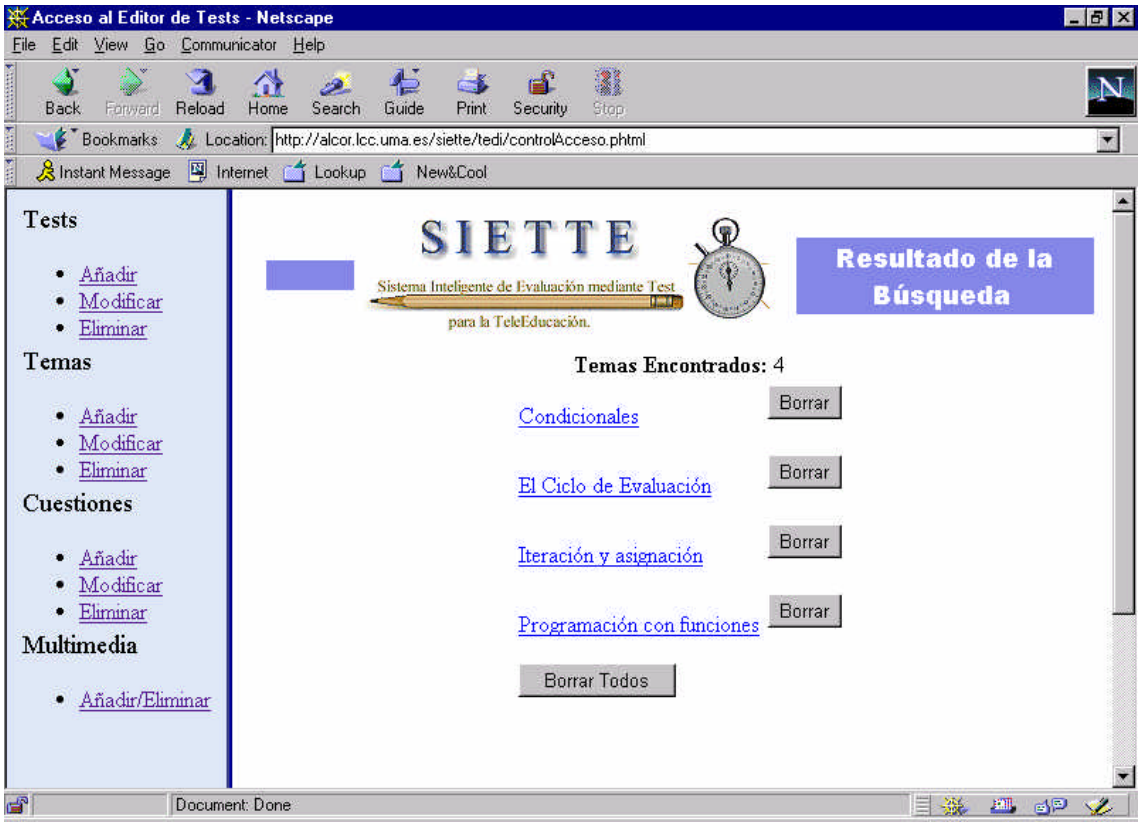


Figura 29. Resultado de la búsqueda tras seleccionar la opción del menú: “Eliminar Tema”.

2) Elegir el tema deseado y ver cuáles son sus datos. Si tras ver los datos del tema (en modo de sólo lectura) se desea eliminar dicho tema, basta con pulsar el botón ***Eliminar Tema*** que aparece al final de los datos del tema.

7.2.6. SECCIÓN CUESTIONES.

Una cuestión en el sistema SIETTE estará definida por una serie de propiedades generales a partir de las cuales se podrá identificar y calibrar dicha cuestión entre otras cuestiones existentes en la asignatura. Los datos aquí suministrados contribuyen a la elección de una pregunta frente a otras a la hora de plantear un test a un alumno. Una cuestión estará formada básicamente por el enunciado de la pregunta, la respuesta correcta y por las posibles respuestas incorrectas (debemos recordar que los tests a crear son tests del tipo “verdadero-falso”).

Tres son las operaciones que el sistema de edición permite realizar sobre las cuestiones de la base de conocimiento de una asignatura:

(a) Creación de una nueva cuestión.

(b) Modificación de los datos de una cuestión previamente creada.

(c) Eliminación de los datos de una cuestión.

7.2.6.1. Creación de una nueva cuestión.

Para añadir una cuestión al banco de cuestiones de una asignatura sólo tiene que hacer clic con el ratón sobre el enlace del menú del editor de la sección “Cuestiones”, titulado “Añadir”. A continuación, se mostrará en su navegador una página HTML solicitando los datos que definen a una cuestión. Estos datos son los datos generales de la cuestión, los datos relativos a la pregunta y los datos relativos de las posibles respuestas.

♦ **Datos generales de la cuestión.** Son todos aquellos datos que permiten identificar y calibrar las preguntas del test (ver figura 30). Estos datos son:

Título: Nombre que identifica la cuestión entre aquellas que forman la misma asignatura. No se permitirá repetir nombres de cuestiones y no es obligatorio dar un nombre a la cuestión ya que en este caso es el sistema el que le asigna un nombre único.

Grado de dificultad: Indicador que dice cómo de difícil es la cuestión dentro del tema del que forma parte. Es uno de los tres parámetros que sirven para calibrar las preguntas y permitir identificar la más apropiada para cada alumno. Los otros dos son calculados implícitamente por el sistema, y son: el grado de adivinanza y el índice de discriminación.

Número de alternativas a mostrar: Definen el número de respuestas *incorrectas* asociadas a la pregunta del test que se está creando. Un profesor puede definir el número de alternativas que desee y con esta opción puede limitar el número de las que se van a ser mostradas a los alumnos en el generador de tests. Si este número es inferior al número de respuestas incorrectas definidas, el sistema elegirá aleatoriamente el número de alternativas aquí especificado. Esto impide aún más el repetir la misma pregunta incluso a usuarios diferentes.

Temas a los que pertenecerá la cuestión: En este apartado se comunica al sistema los temas a los que pertenece la cuestión que se está creando. Basta con seleccionar aquellas que se deseen, para establecer la relación: “ cuestión – temas”.

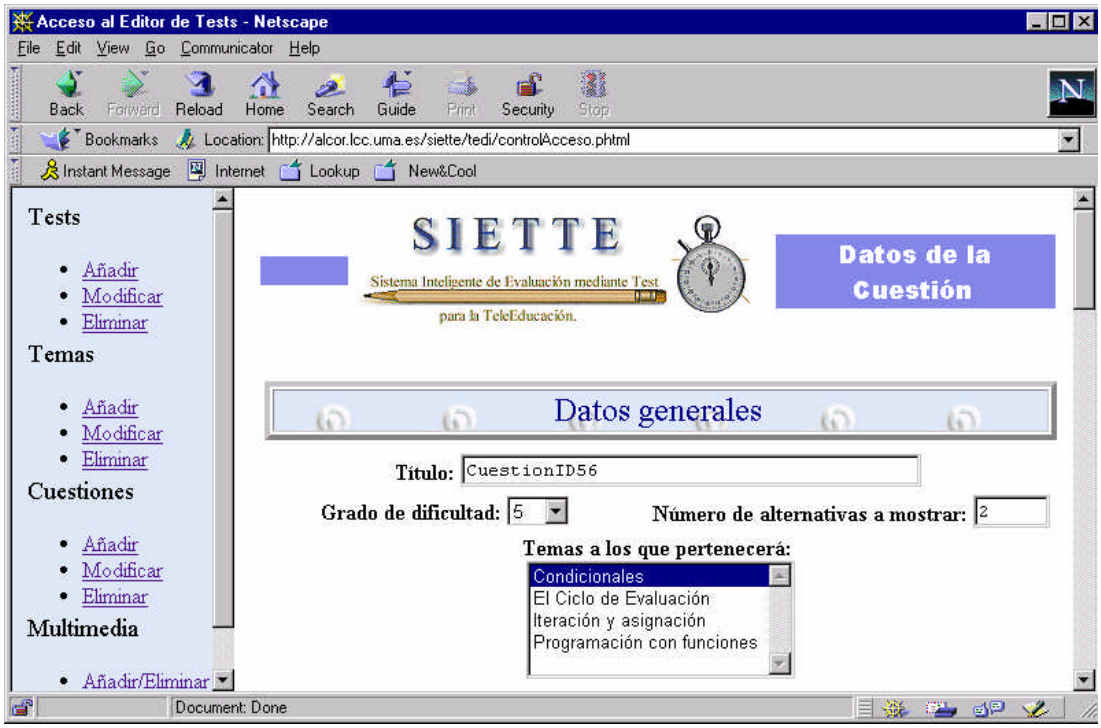


Figura 30. Datos generales de una cuestión del test.

♦ **Datos relativos a la pregunta del test.** La pregunta del test vendrá identificada por el enunciado y por algún tipo de ayuda (ver figura 31).

Enunciado: En este campo debe escribir el texto que aparecerá como enunciado de la pregunta en el test. El texto que puede redactar es texto normal o bien texto HTML para dar el formato al enunciado que se mostrará al alumno.

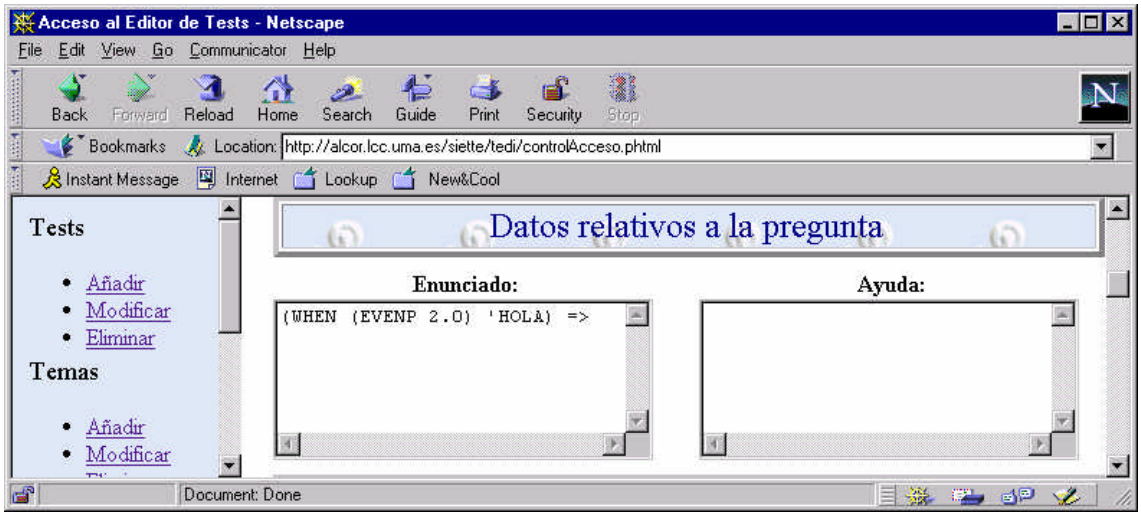


Figura 31. Datos de la pregunta. Enunciado y ayuda.

Por tanto, el formato y la visualización de la pregunta serán totalmente flexibles.

Si lo que desea es crear un ***esquema de pregunta***, es decir, que el enunciado no sea estático sino que sea dinámico (una plantilla del enunciado que se generará cuando se plantee este pregunta al alumno), también puede escribir en este campo código PHP. Dicho código será ejecutado si esta cuestión es seleccionada por el generador de test, para ser planteada al alumno. La mayor utilidad de esta opción se alcanza cuando se une la potencia de la WWW, con el lenguaje PHP y con las bases de datos que permiten su acceso a través de la WWW (para más información acceder a los manuales disponibles en http://php.iquest.net y http://www.postgreSQL.org). Básicamente, se trata de introducir una plantilla del enunciado que puede generar diferentes instancias para ese enunciado. Esto dará más aleatoriedad en el test que se generará para cada alumno, siendo poco posible repetir la misma instancia del esquema a distintos alumnos.

Ayuda: Coloque aquí cualquier aclaración o explicación de la pregunta que facilite su solución por parte del alumno. Si el test al que pertenece la pregunta es de autoevaluación y se aporta información en este campo, el alumno verá al lado de la pregunta del test un botón de ayuda. Al pulsar dicho botón se mostrará el texto tal y como aquí se redacte. En cambio, si el test es de evaluación, no se mostrará la ayuda al alumno aunque ésta exista.

♦ **Datos relativos a las posibles respuestas.** Los datos relativos a las respuestas de la pregunta son: el modo en que se desean visualizar las preguntas, la respuesta correcta y las respuestas incorrectas asociadas con dicha pregunta.

El modo de visualización de las respuestas de la pregunta dentro de un test será el mostrado en la figura 32.

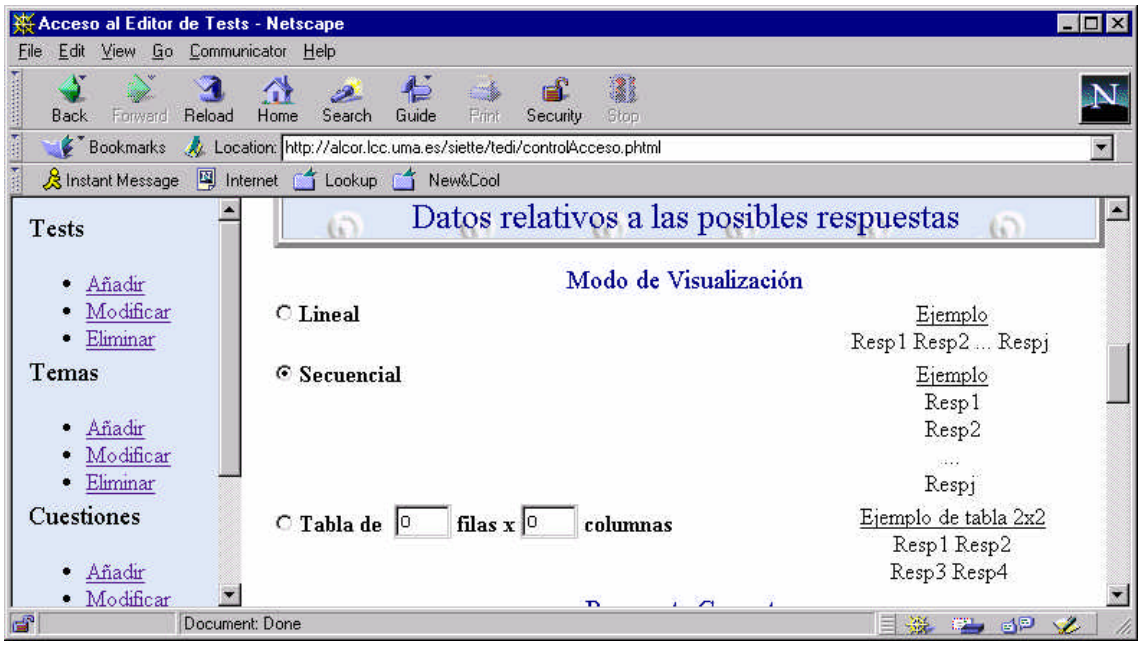


Figura 32. Modo de visualización de las respuestas de la pregunta dentro del test.

Respuesta correcta y respuestas incorrectas: En este apartado se deben suministrar el texto y alguna explicación sobre las respuestas a la pregunta (ver figura 33).

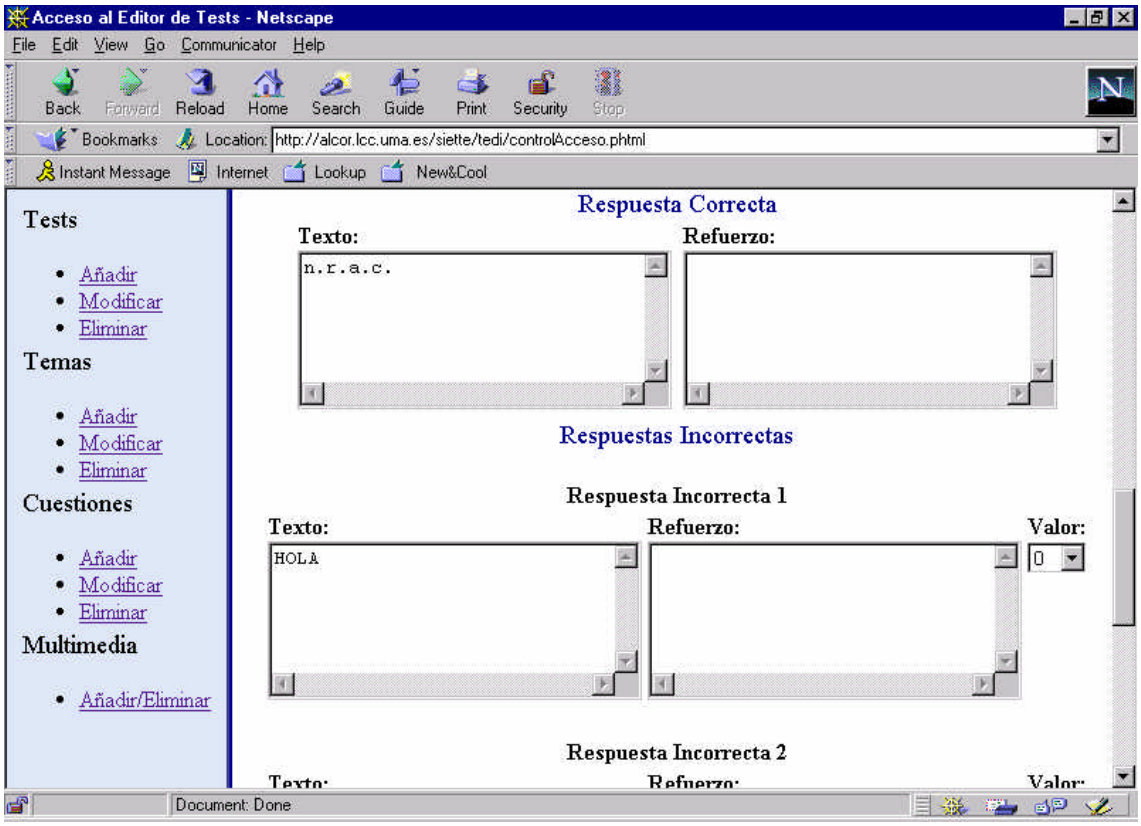


Figura 33. Respuesta correcta e incorrectas de una pregunta de test

Al igual que en el campo de texto del enunciado de la pregunta, en el campo de la respuesta puede colocar texto normal, texto HTML para dar formato, y texto con código PHP si quiere crear plantillas de respuestas en lugar de dar las respuestas directamente. El texto que coloque en el refuerzo (aclaraciones, explicaciones, etc.) sólo se mostrará en caso de que la pregunta que está creando forme parte de un test de autoevaluación. En caso contrario no se mostrará aún cuando se haya suministrado tal información.

♦ **Crear la cuestión.** Para acabar de crear la cuestión, el usuario puede encontrarse con los botones que se muestran en la figura 34.

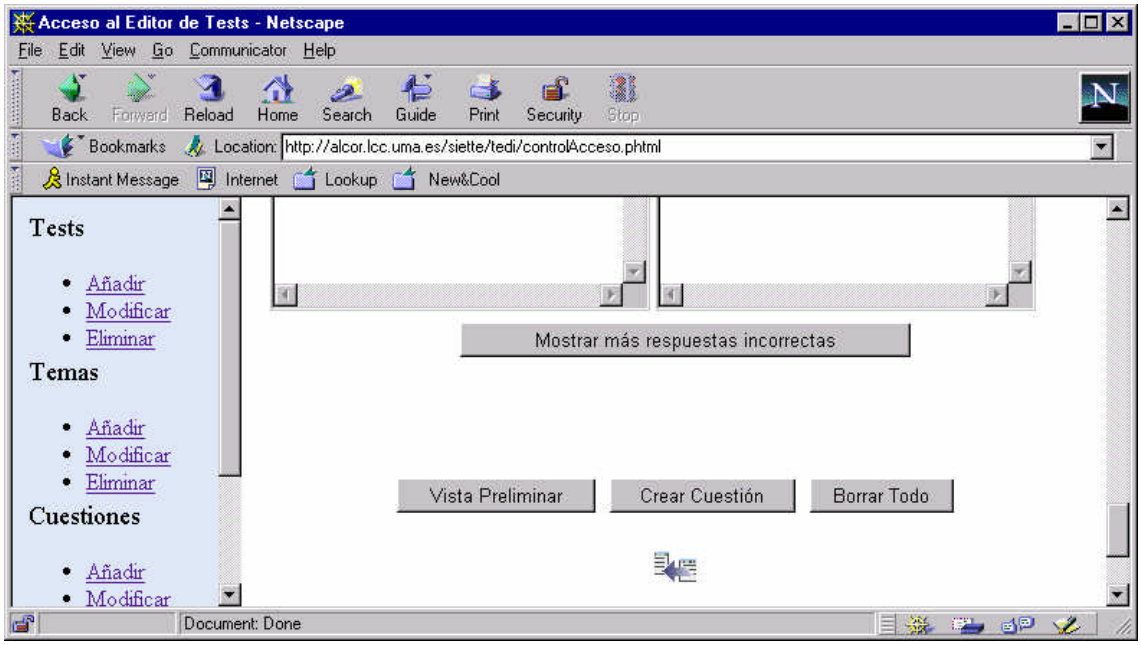


Figura 34. Operaciones que se pueden realizar al crear la cuestión.

Mostrar más respuestas incorrectas: Inicialmente, se muestran tres campos de texto para que se den tres posibles respuestas incorrectas para la pregunta. En caso de querer asociar a la pregunta más de tres respuestas incorrectas, se debe pulsar el botón etiquetado como ***Mostrar más respuestas incorrectas,*** el resultado será el mismo formulario HTML que teníamos en este punto sólo que se han añadido al final tres nuevos campos relativos a respuestas incorrectas.

Vista preliminar: Si se pulsa este botón, se mostrará la disposición y formato que tendrá la pregunta al ser planteada al alumno. A su vez, si existe alguna plantilla de enunciado en la cuestión (bien de la pregunta, bien de la respuesta), se mostrará el resultado de instanciar dicha plantilla o bien, los errores de compilación (sí los hay).

A pesar de que en la vista preliminar de la cuestión siempre se muestra en primer lugar la respuesta correcta y seguidamente las respuestas incorrectas, cuando esta pregunta sea planteada por el generador de tests al alumno, el orden de éstas variará.

Si no se producen errores al pulsar el botón ***Crear Cuestión*** se almacenarán los datos suministrados en el servidor.

7.2.6.2. Modificación de los datos de una cuestión previamente creada.

Antes de modificar la cuestión, debemos identificar qué cuestión queremos cambiar. Para ello, haga clic en el enlace “Modificar” de la sección “Cuestiones” del menú del editor. Le aparecerá el siguiente formulario de búsqueda de cuestiones. Ver figuras 35 y 36.

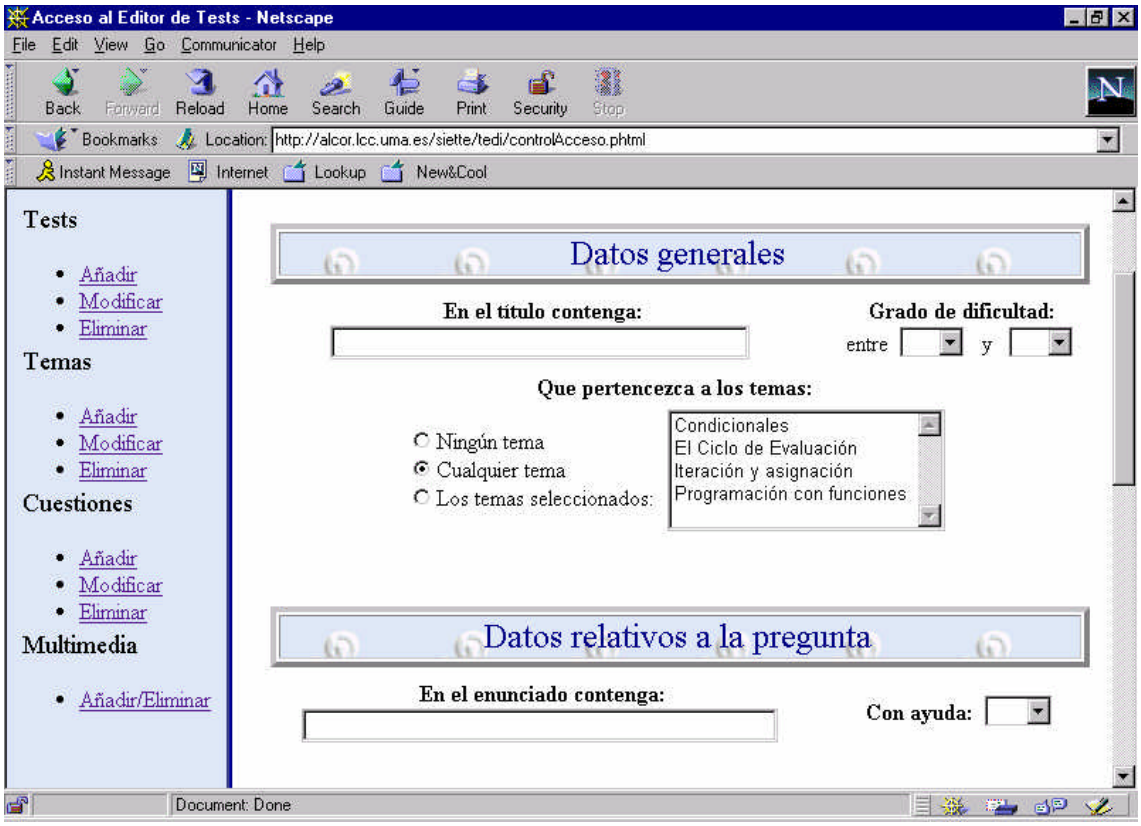


Figura 35. Formulario de búsqueda de cuestiones. Datos generales y enunciado de la pregunta.

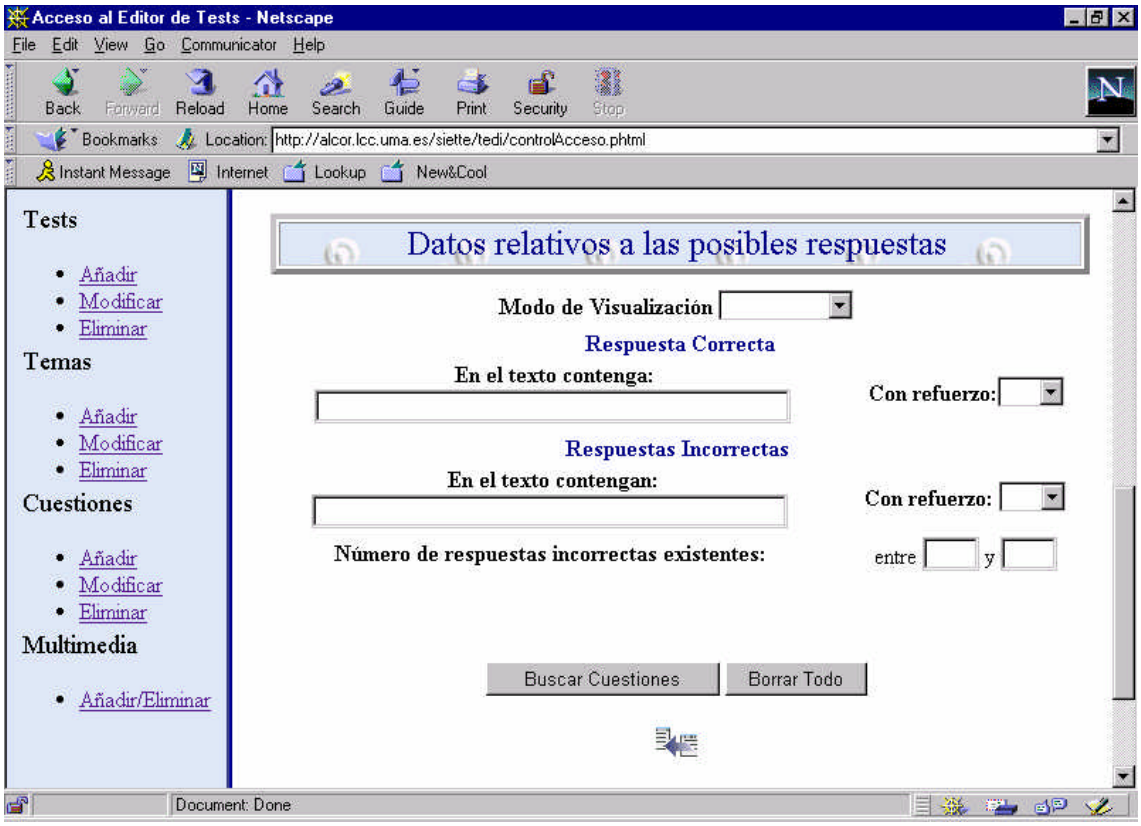


Figura 36. Formulario de búsqueda de cuestiones. Datos de las posibles respuestas.

A través de estos formularios podrá restringir la búsqueda de la cuestión o las cuestiones que desea modificar. Después de pulsar el botón ***Buscar Cuestiones***, aparecerá una pagina HTML como la mostrada en la figura 37. Es decir, la lista de las cuestiones encontradas. Haciendo clic en cualquiera de ellas podrá ver de nuevo sus datos en ventanas como las ya mostradas en las figuras de la 30 a la 34. En este caso, en la figura 34 en lugar de aparecer el botón "Crear Cuestión" aparecerá el botón ***Modificar Cuestión***.

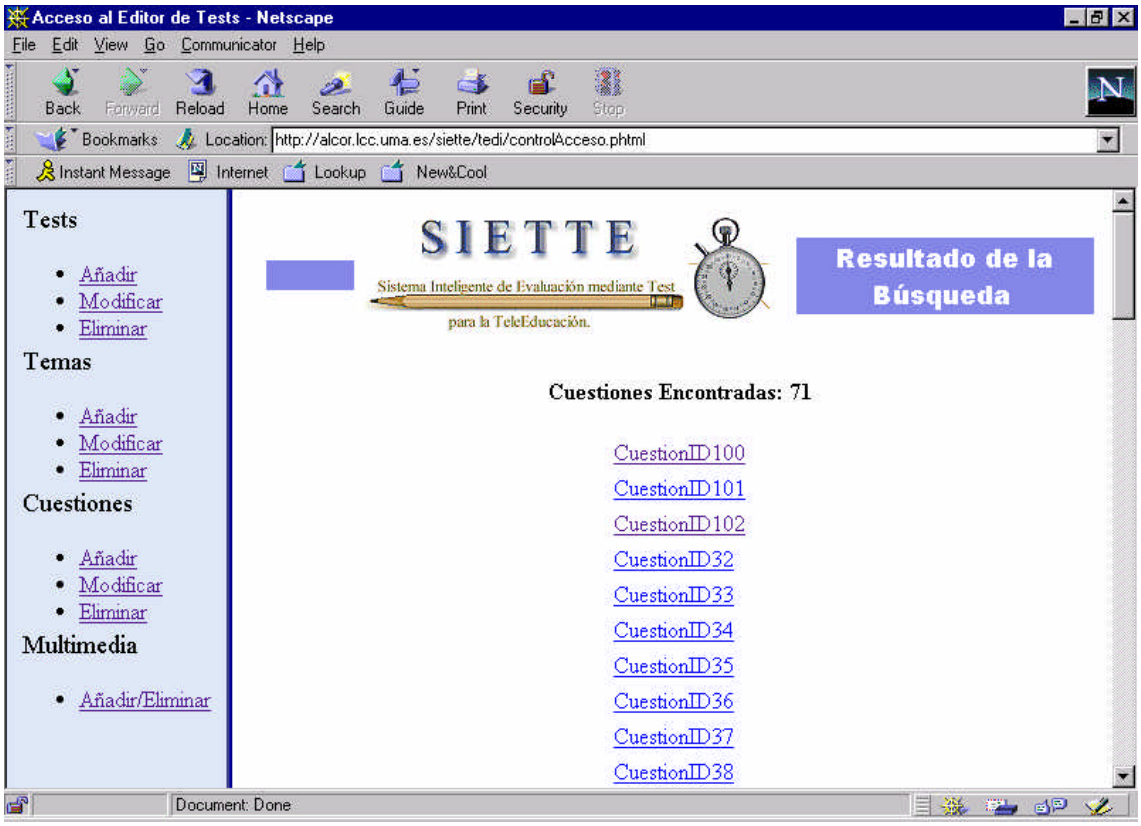


Figura 37. Resultado de la búsqueda de cuestiones de una asignatura.

7.2.6.3. Eliminación de los datos de una cuestión previamente creada.

Al eliminar los datos de una cuestión se eliminarán los datos suministrados en los formularios anteriores así como las relaciones establecidas entre la cuestión a eliminar y los temas que las poseen.

Para proceder al borrado de una cuestión sólo tiene que hacer clic sobre el enlace “Eliminar” de la sección “Cuestiones” del menú. De nuevo le aparecerá el formulario de búsqueda de las figuras 35 y 36 y como resultado de la búsqueda una página HTML como la mostrada a continuación (figura 38). A partir de aquí podrá borrar las cuestiones que desee o bien, hacer clic sobre aquella de la que quiera saber más datos antes de borrarla.

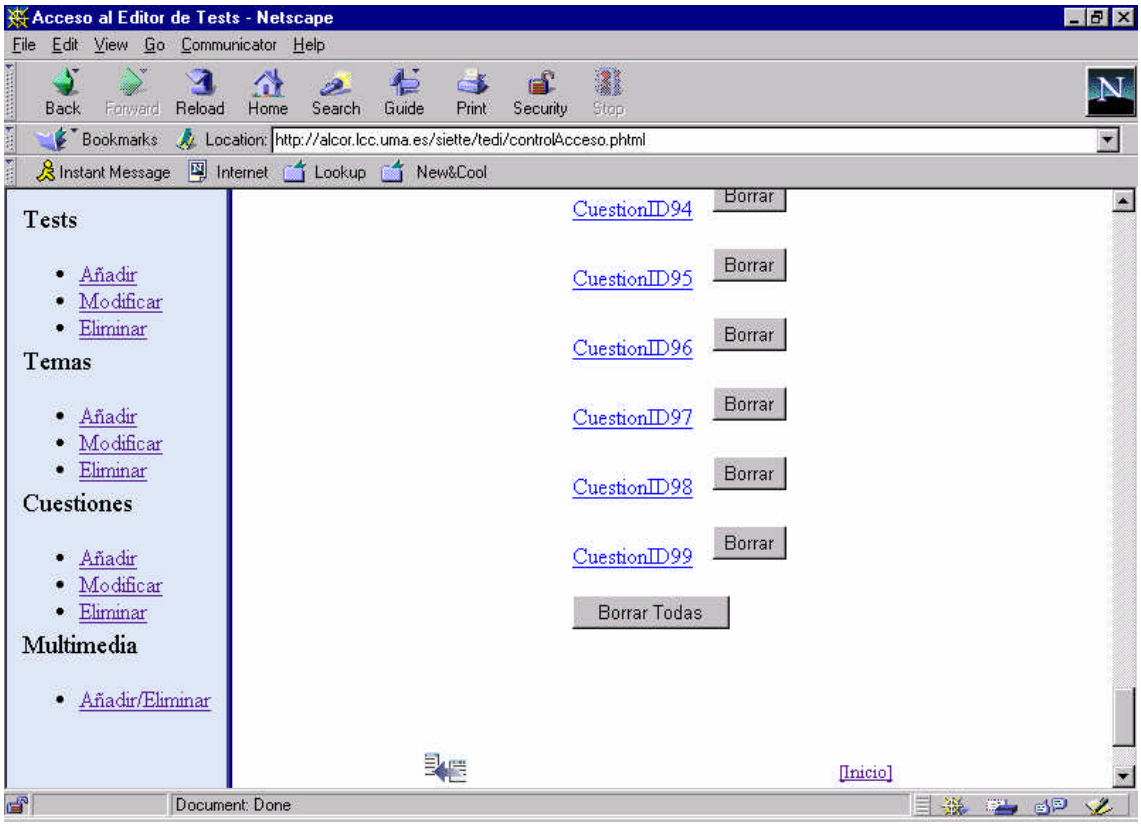


Figura 38. Resultado de la búsqueda para llevar a cabo la eliminación de cuestiones.

7.2.7. SECCIÓN MULTIMEDIA.

Para aumentar la experiencia del aprendizaje de los alumnos e ilustrarles escenarios del mundo real, se pueden integrar componentes multimedia en los tests. Si el profesor quisiera evaluar a sus alumnos sobre géneros musicales por ejemplo, podría incluir ficheros de audio en sus tests, para ilustrar diferentes géneros musicales. Del mismo modo, podría evaluar a los alumnos sobre especies vegetales, incluyendo ficheros con imágenes como parte de los enunciados de las preguntas y/o de las posibles respuestas que aparecen en el test. En definitiva, un test podrá contener cualquier tipo de ficheros multimedia si el profesor así lo desea.

Para añadir el contenido multimedia en un test debe seguir los siguientes pasos:

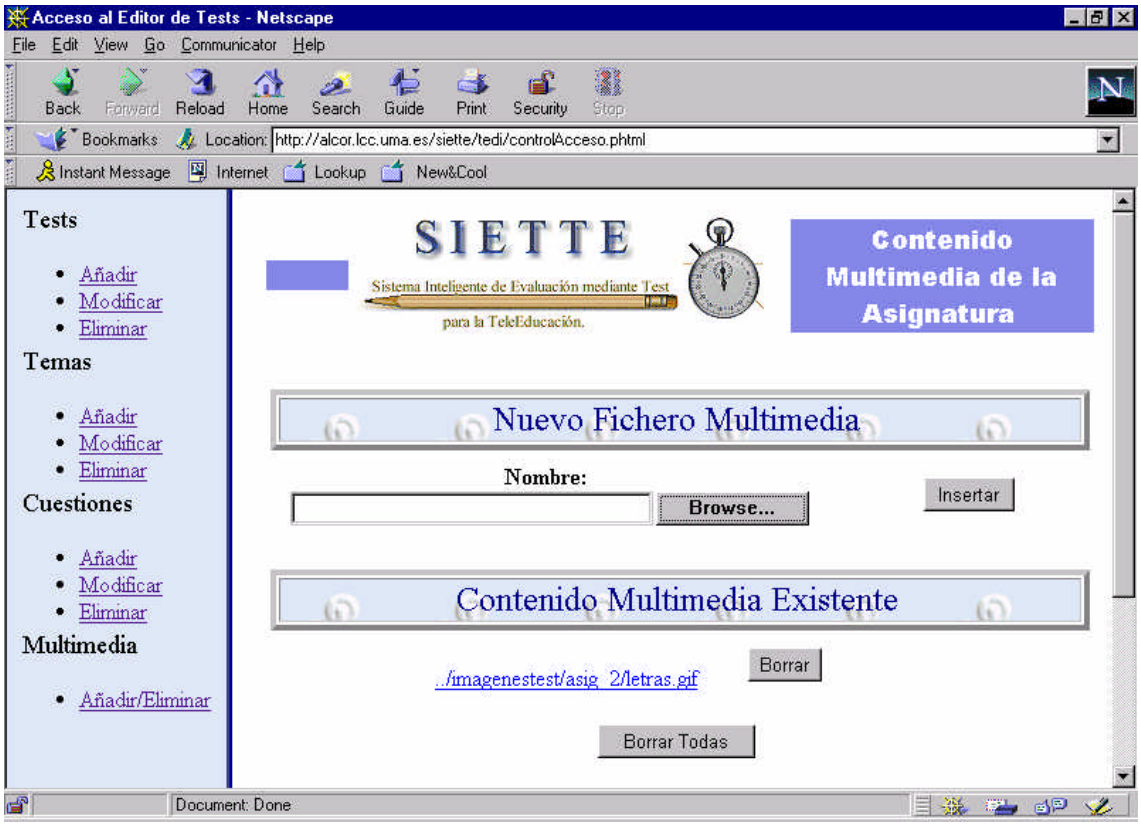


Figura 39. Añadir/Eliminar contenido multimedia a una asignatura.

Para insertar contenido multimedia debe hacer clic sobre el botón ***Browse***. A continuación, se abrirá una ventana con el sistema de ficheros del disco de su PC (ver figura 40).

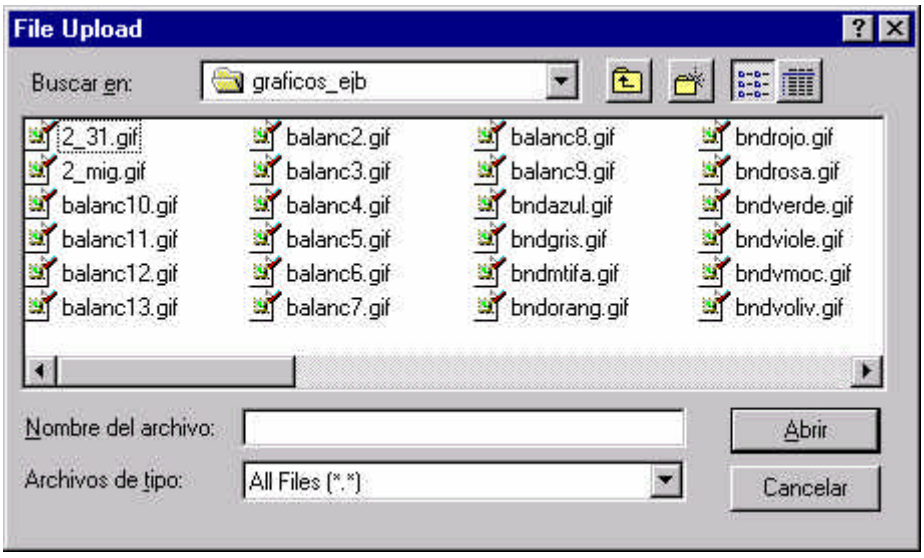


Figura 40. Sistema de ficheros del disco local.

Seleccione el fichero multimedia que desee y pulse ***Abrir***. Esta ventana se cerrará y volverá a visualizar la página HTML de la figura 39. Ahora aparecerá relleno el campo **Nombre** con el nombre del fichero elegido.

Seguidamente, pulse el botón ***Insertar*** y el fichero dado se copiará y almacenará en el servidor. Si no hubo errores, la próxima vez que visualice la página de la figura 39 aparecerá en la sección de **Contenido Multimedia Existente** un enlace al fichero almacenado en el servidor WWW. Pulsando este enlace podrá identificar dicho fichero. Además, el enlace mostrado es el que deberá colocar en los enunciados de las preguntas y/o de las posibles respuestas si quiere que al generar la pregunta en el test, ésta se visualice correctamente.

Finalmente, para borrar algún fichero sólo debe pulsar los botones de **borrado**. El sistema pedirá confirmación antes de realizar el borrado del fichero o de los ficheros dados.

7.3. GUÍA PARA EL ALUMNO.

El alumno interactuará directamente con el generador de tests, para evaluarse y/o autoevaluarse dependiendo del test que elija. Por tanto en esta sección explicaremos qué tipo de pantallas se encontrará un alumno a lo largo de un test y qué datos se verán en cada una de ellas.

El alumno debe seguir los siguientes pasos:

1) Ejecutar el navegador instalado en su PC (ver apartado “Requerimientos Software” para más información).

2) Conectarse a la URL donde está instalada la aplicación (ver apartado “Instalación del sistema SIETTE”).

3) A continuación, se mostrará la página principal de la aplicación (ver figura 16) deberá hacer clic sobre el enlace ***Aula de Tests***.

4) Identificación del alumno.

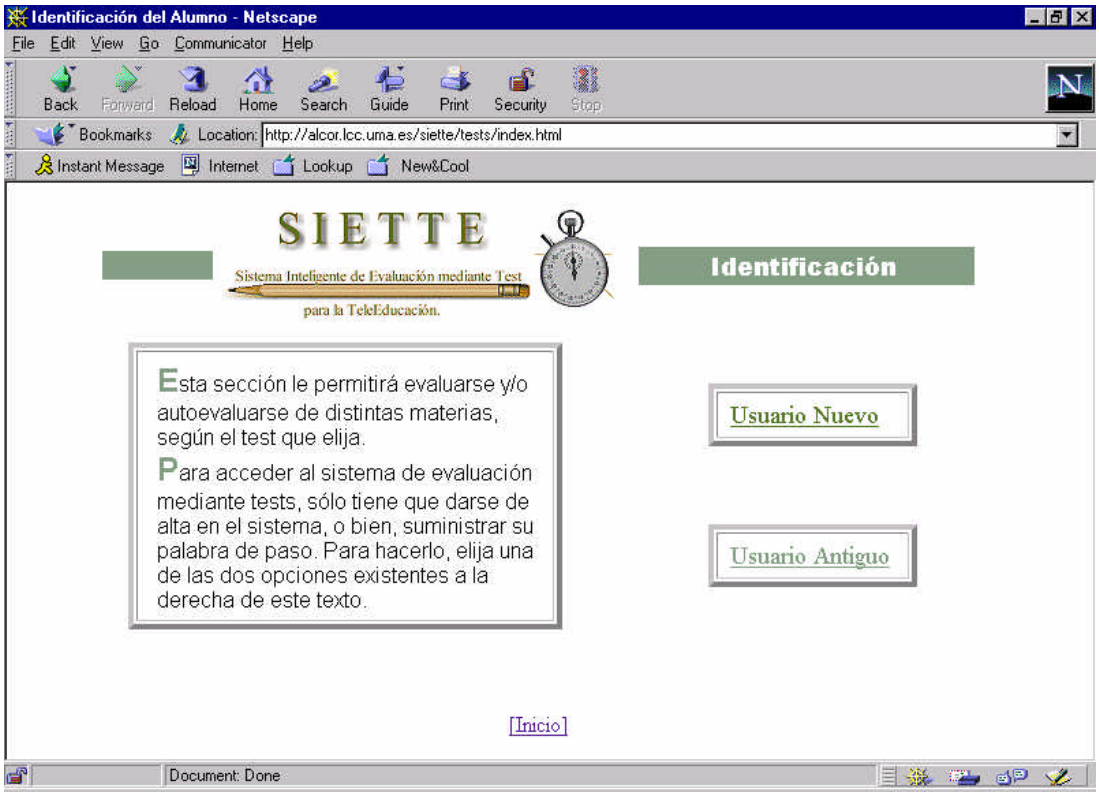


Figura 41. Identificación del alumno.

Independientemente de sí el alumno está o no registrado en el sistema, se le pedirá una clave y código de acceso, al mismo tiempo que se le muestran los tests disponibles en ese momento junto con sus descripciones. Ve figura 42.

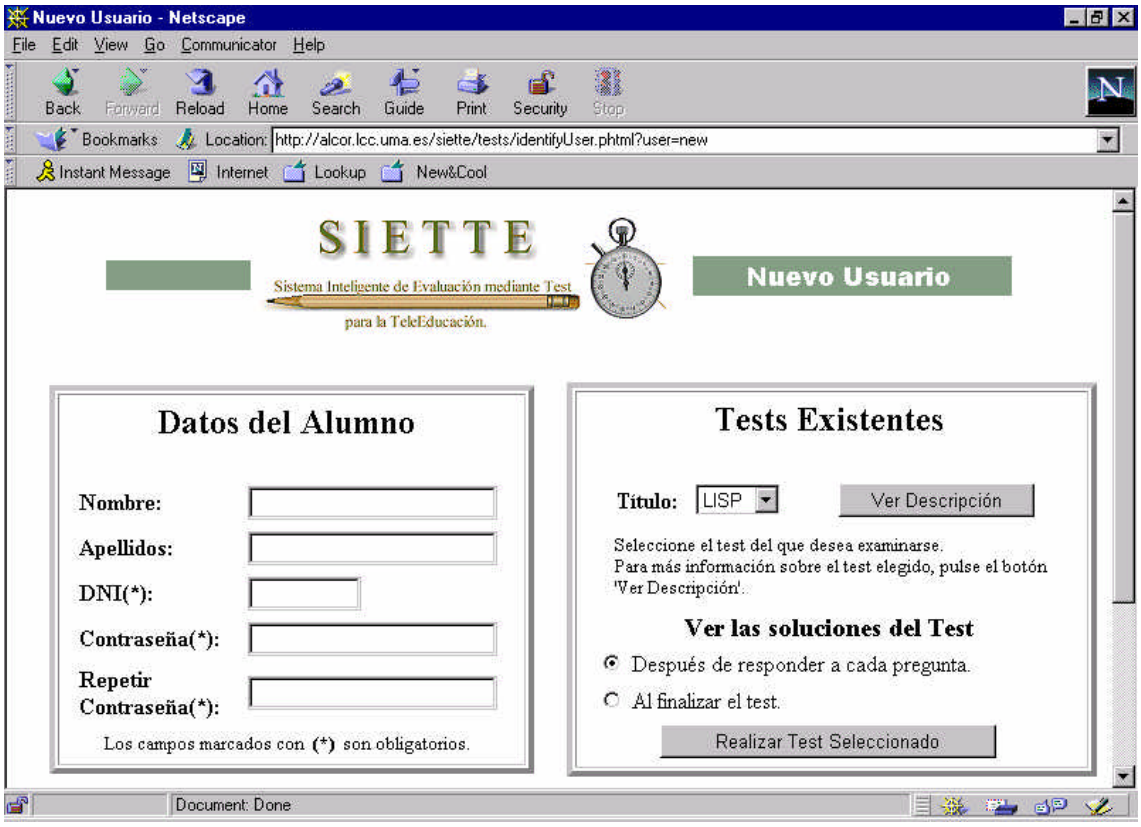


Figura 42. Tests existentes.

Pulsando el botón ***Ver Descripción*** se abrirá una nueva ventana en la que se mostrará la descripción del test seleccionado.

5) Elegir el test que se desee y pulsar el botón ***Realizar Test Seleccionado.***

6) En este momento, se mostrará una nueva página HTML en la que se mostrarán algunos datos del test como el nombre, periodo máximo de disponibilidad de dicho test a los usuarios, descripción del test y avisos relativos al tipo de test y nivel mínimo exigido para superar el test.

7) Pulsar el botón ***Comenzar Test***. Tras esta operación, el sistema empezará a plantear preguntas al alumno así como las estimaciones sobre el conocimiento que el sistema va calculando a lo largo del test (ver figuras 43 y 44).

8) Al ser preguntas del tipo "verdadero-falso", se puede: no contestar y seguir el test (la pregunta será evaluada como si se hubiese contestado incorrectamente), o bien seleccionar una de las respuestas mostradas.

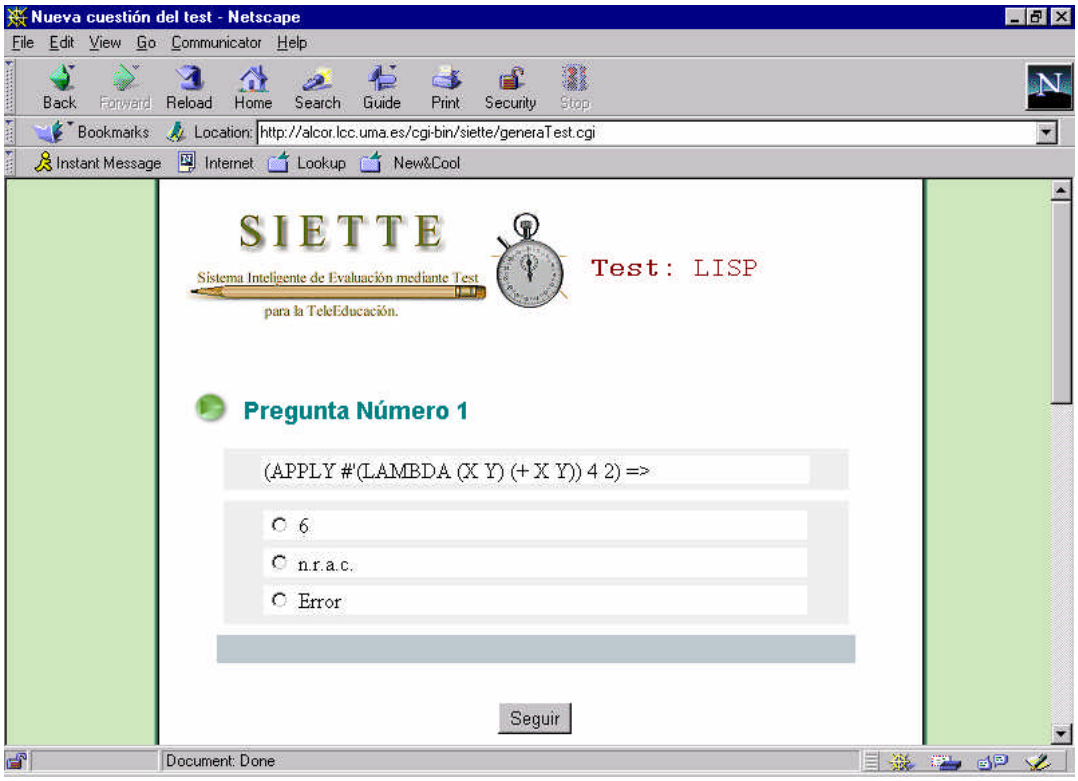


Figura 43. Formato de las preguntas de un test en SIETTE.

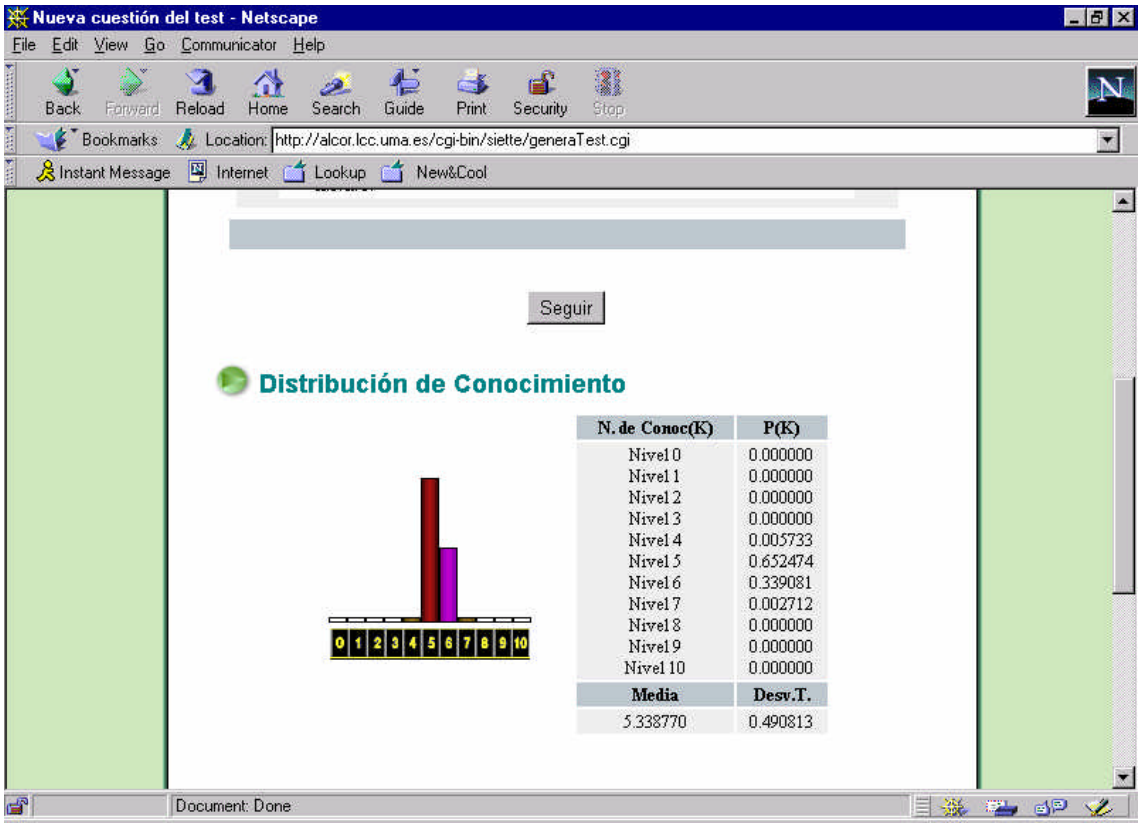


Figura 44: Estimación del nivel de conocimiento de un alumno en un momento de la sesión de test.

9) Una vez pensada la respuesta a la pregunta, se ha de pulsar el botón ***Seguir***. Entonces pueden ocurrir las siguientes cosas:

♦ Si antes de comenzar el test se eligió ver las respuestas después de responder a cada pregunta, se mostrará la corrección de la pregunta que se acaba de responder(ver figura 45).

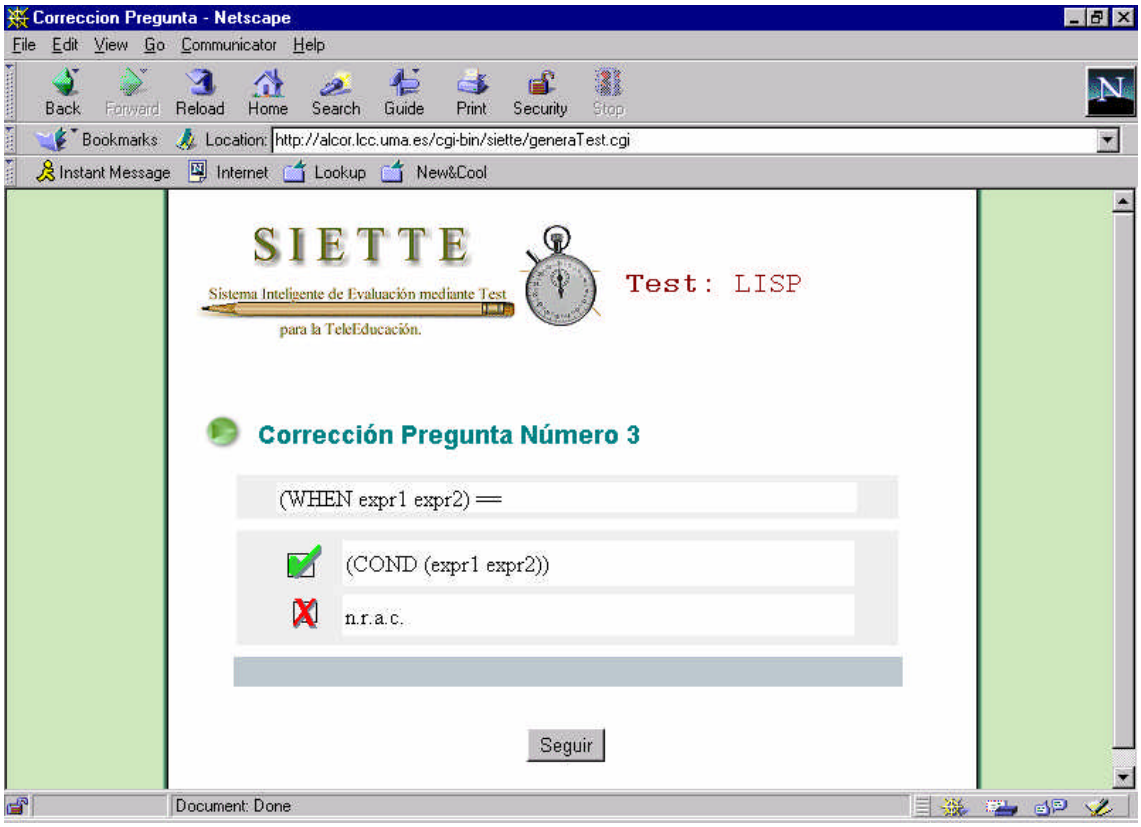


Figura 45. Corrección a una pregunta del test. Alumno contestó incorrectamente.

♦ En caso contrario, el sistema planteará una nueva pregunta junto con la distribución de conocimiento calculada hasta ese momento.

10) El usuario podría intentar hacer trampas y responder de nuevo una pregunta anteriormente contestada, y de la cual ha visto la respuesta correcta. En este caso, el sistema SIETTE detecta el intento de falsear el test y le vuelve a plantear la pregunta que le correspondía contestar, dándole un aviso de que no puede alterar el orden del test. Ver figura 46.

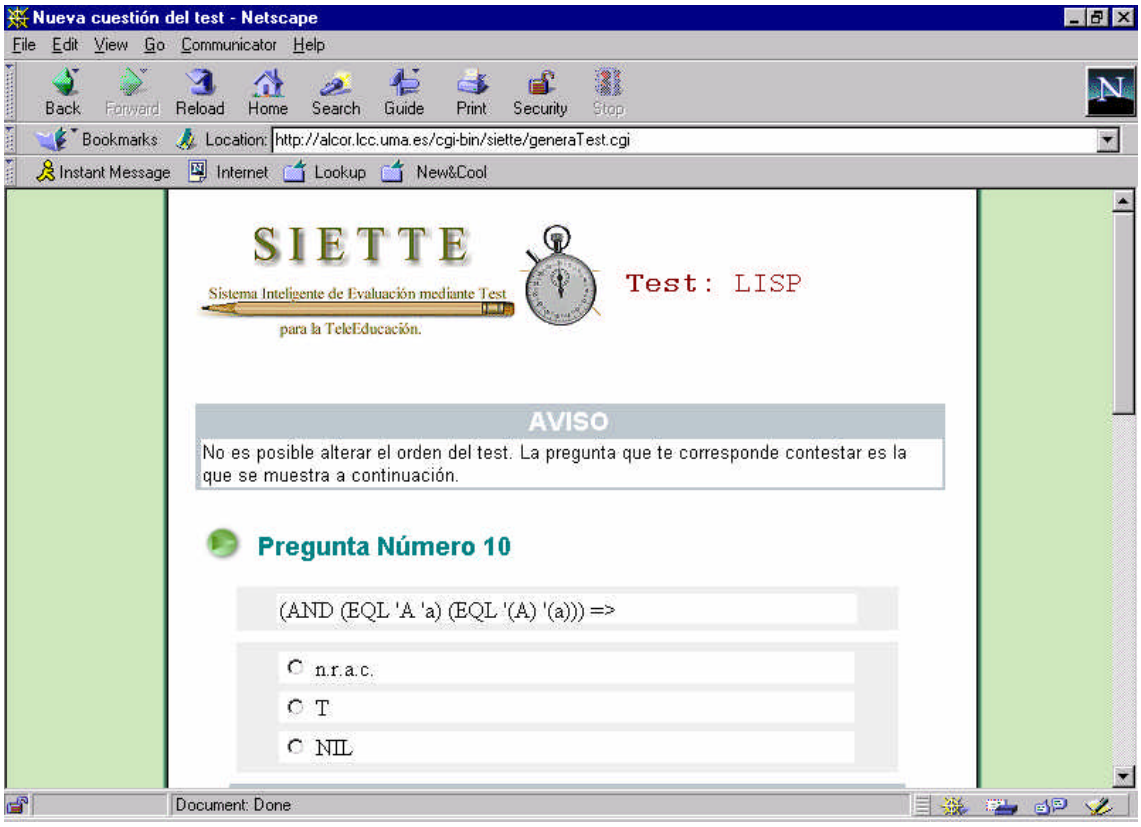


Figura 46. Aviso mostrado por el sistema al intentar alterar el orden del test.

11) Continuar respondiendo preguntas hasta que el sistema muestre la calificación final (ver figuras 47 y 48). El modo de interpretar la calificación es la siguiente:

➢ El test que se plantea es un test adaptativo lo que indica que el sistema intentará calificar al alumno en “apto” o “no apto”. Una vez que logra clasificar al alumno en una de estas dos categorías el test finaliza dando uno de estos dos resultados: APROBADO o SUSPENSO. Por tanto, no se trata de responder un número fijo de preguntas, en función del cual se califica al alumno sino que según estén calibradas las preguntas unas puntuarán más que otras y unas veces será necesario responder a mayor número de preguntas que otras veces. Si el alumno desea variar la estimación calculada por el sistema (puede lograr que el sistema incremente el nivel estimado de su conocimiento o lo disminuya) sólo tiene que volver a realizar el test (se permite si el test es de autoevaluación). De este modo, las sucesivas veces que el alumno realice el mismo test, el sistema intentará plantear al alumno la pregunta más adecuada para el conocimiento estimado en la última sesión.

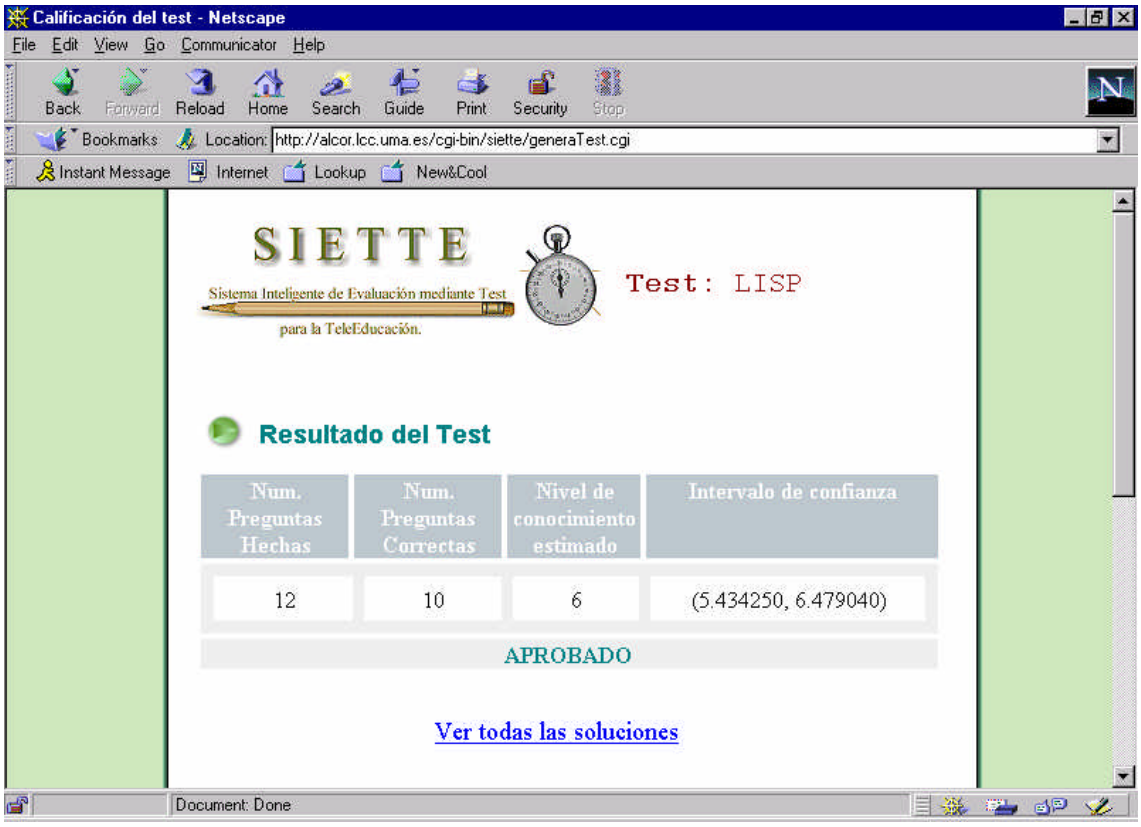


Figura 47. Calificación final del test.

Haciendo clic sobre el enlace ***Ver todas las soluciones*** pueden obtenerse los enunciados de las preguntas planteadas y la corrección a las respuestas dadas a tales preguntas.

El intervalo de confianza mostrado es el intervalo en el que estará la estimación realizada por el sistema para tener un grado de credibilidad del 95% o mayor.

La distribución final alcanzada está del tipo:

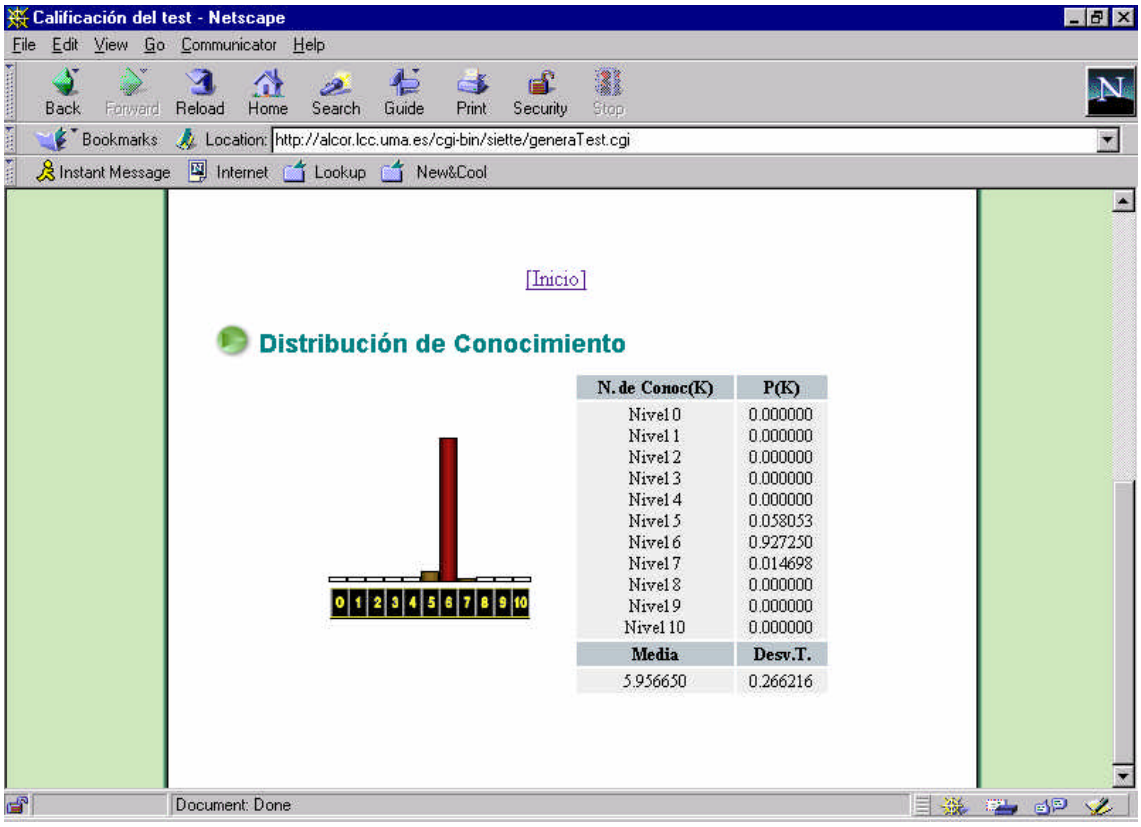


Figura 48. Distribución de conocimiento de un alumno al finalizar el test.

12) Iniciar una nueva sesión de test, si se desea. Para ello vuelva a la página inicial de SIETTE.

8. CONCLUSIONES Y LÍNEAS FUTURAS

La mayor dificultad a la hora de implementar muchos de los algoritmos citados a lo largo de esta memoria, es la necesidad de realizar un gran estudio empírico para calibrar y construir el banco de preguntas para los posibles tests, antes de poder usar dichos algoritmos (la mayoría de ellos basados en la teoría IRT). Tal estudio es raramente abordable por pequeñas organizaciones que ofrecen entrenamiento a sus alumnos.

Otras de las dificultades que surgen al usar dichos algoritmos es el problema del contenido que abarcan los tests generados. La mayoría de ellos emplean estrategias de selección de preguntas ajenas a las áreas de contenido del currículum de la materia a evaluar, y a la que pueden pertenecer dichas preguntas. Por tanto, los tests generados podrían no cubrir los objetivos de la instrucción, ya que importantes áreas podrían no tratarse y generar así una valoración del alumno menos precisa.

A su vez, las primeras características que resaltan del uso de la WWW en los entornos educacionales son: la posibilidad de tener documentos multimedia, capacidad hipertexto y arquitectura cliente/servidor que permite la enseñanza a distancia.

El objetivo del sistema SIETTE ha sido el desarrollar un sistema que proporcione solución a los problemas que presentan los tests adaptativos, al tiempo que une el dinamismo de tales tests con las características que ofrece la Web como entorno de aprendizaje y por tanto, como medio de evaluación de los conocimientos previamente impartidos.

Para la realización de dicho sistema, se ha usado el razonamiento basado en probabilidades, llegándose a la conclusión de que el uso eficiente y completo de tal tipo de razonamiento radica en:

- las definiciones de las variables que capturan los elementos destacados de cada test (conceptos de la instrucción que hay que evaluar),
- la estructura de la distribución de probabilidades, y
- en la independencia condicional que capturan las relaciones más importantes entre los elementos implicados en un test adaptativo (el nivel estimado de conocimiento que

tiene el alumno y las respuestas asociadas a cada una de las preguntas existentes en el banco de preguntas).

Por tanto, si las relaciones necesarias para el razonamiento deductivo (proporcionalidad de respuestas correctas dado un nivel de conocimiento) y las creencias a priori sobre el parámetro desconocido (nivel de conocimiento del alumno) pueden ser expresadas como relaciones en una red de probabilidades matemáticas, entonces el Teorema de Bayes puede proporcionar un principio de razonamiento inductivo que justifica la precisión de las estimaciones del nivel de conocimiento, dentro de dicha red.

Llegado a este punto, pronosticaremos algunas **directrices futuras** que podrían aplicarse sobre el sistema de tests adaptativos implementado en el sistema SIETTE, y que lo mejorarían sustancialmente:

- En primer lugar, implementar un *procedimiento de aprendizaje* que actualice los valores inicialmente asignados a los parámetros que caracterizan a cada pregunta, combinando ese valor inicial con la información histórica de cada alumno que ha interactuado con el sistema.

Un modo de dotar al sistema de cierta habilidad de aprendizaje sería hacer que el nivel de dificultad asignado inicialmente por el profesor a cada pregunta, variase en función de la proporción de alumnos que, a pesar de estar catalogados en poseer un cierto nivel de conocimiento, responden a dicha pregunta correcta o incorrectamente.

Igualmente, podrían generarse perfiles de aprendizaje de los alumnos con los datos calculados sobre la base de conocimiento del sistema. Dichos datos servirían a los profesores para afrontar o modificar la instrucción que realizan, o bien cambiar las especificaciones dadas en los tests que crearon.

- Cuando la base de conocimiento del sistema SIETTE creciese suficientemente (existencia de aproximadamente unas 500 preguntas por asignatura), debería de mejorarse el algoritmo de selección de preguntas para que éste fuese más eficiente, ya que en los tests adaptativos hay que esperar a conocer la respuesta que el alumno da a la pregunta actual, para poder plantearle la siguiente pregunta más adecuada a su nivel de conocimiento.

Un primer procedimiento de búsqueda eficiente de la siguiente pregunta, podría consistir en *realizar dos búsquedas mientras el examinando responde la pregunta*

actual. Una búsqueda seleccionaría la pregunta más informativa para el alumno suponiendo que la pregunta que está contestando actualmente ha sido correcta, mientras que la otra búsqueda seleccionará la pregunta más adecuada suponiendo que la respuesta que dará el alumno es incorrecta. Por tanto, después de que el examinando responda a la pregunta actual, la siguiente pregunta a plantear será alguna de las dos ya buscadas.

En el sistema SIETTE al ser un sistema sobre WWW este método parece ser el indicado, ya que se tiene de por medio el retardo, a veces significativo, de la red para cada una de las preguntas que se le plantean al alumno a lo largo de un test.

- Modelar tests adaptativos del tipo "verdadero-falso" *con más de una respuesta correcta*, que facilitaría sobre todo, la generación de preguntas a partir de plantillas.

Contemplar esto, supondría asociar las curvas características de las cuestiones del test, a las respuestas y no a las preguntas. Es decir, asociada a cada pregunta existe, actualmente en SIETTE, una curva que da la proporcionalidad de responderla correctamente según los niveles de conocimiento definidos (del Nivel 0 al Nivel 10). Y en este caso, habría que asociar a cada respuesta una curva que definiese la proporcionalidad de dar dicha respuesta como correcta según cada nivel conocimiento.

De este modo, el modo de evaluar la corrección de la pregunta se haría del mismo modo que con las preguntas dicotómicas con la salvedad de que la pregunta de n respuestas correctas se dividiría en n preguntas dicotómicas, que serían evaluadas secuencialmente como si todas ellas hubiesen sido planteadas a la vez al alumno y todas ellas hubiesen sido contestadas por éste.

- Por último, pero no por ello menos importante, es la *posible aplicación del sistema SIETTE en un sistema tutorial inteligente* como herramienta de apoyo del módulo de diagnóstico. Una posible forma de utilización sería:

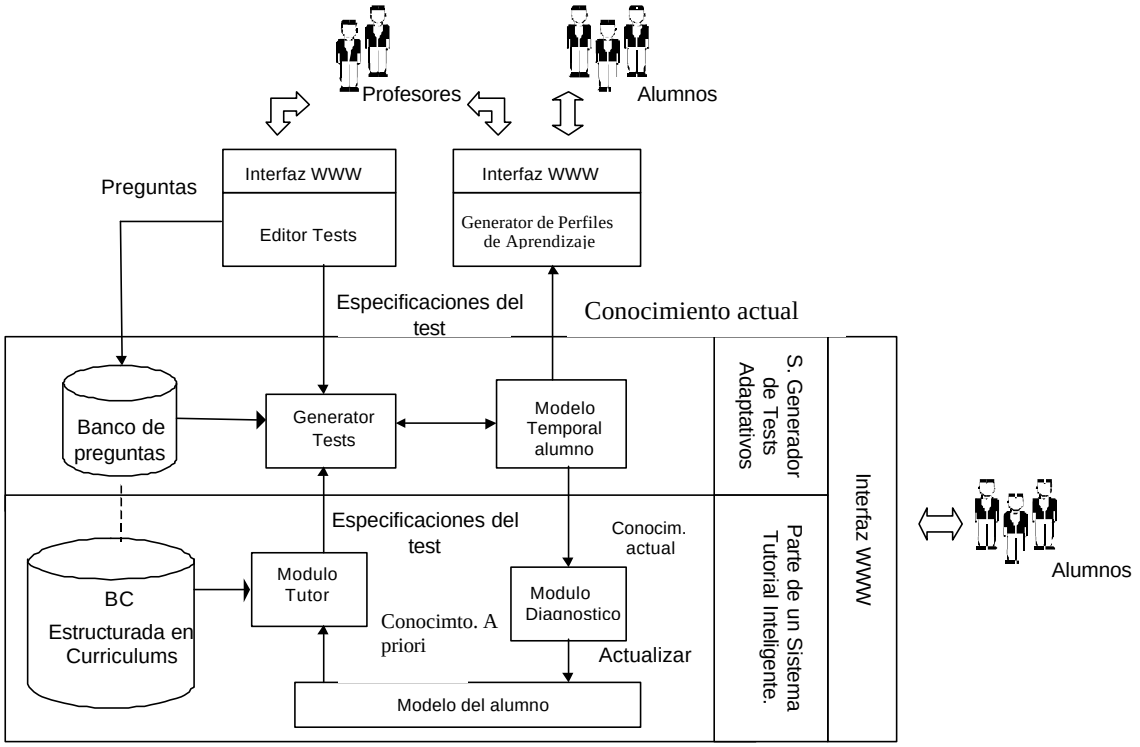


Figura 49. Interacción del Sistema SIETTE con un STI.

Podemos aventurar que los rápidos avances en las líneas de investigación estadísticas incrementaran la utilidad del razonamiento basado en probabilidad en los sistemas tutoriales inteligentes.

APÉNDICE A. BIBLIOGRAFIA Y REFERENCIAS

Birnbaum, A. (1968): Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M. R. Novick, Statical Theories of Mental Test Scores.

Collins, J.A., Greer J.E. y Huang, S.X. (1997): Adaptive Assessment using Granularity Hierarchies and Bayesian Nets

Huang, Sherman X. (1996): On Content-Balanced Adaptive Testing. Computer Aided Learning and Instruction in Science and Engineering, pp. 60-68.

Manual de HTML. URL: http://www.uca.es/manual-html/

Manual de PaintShop Pro. URL: http://www.jasc.com/

Manual de JavaScript. URL: http://developer.netscape.com/docs/manuals/communicator/jsguide4/index.htm

Mislevy, R. (1996). Evidence and Inference in Educational Assessment. CSE Technical Report 414.

Mislevy, R. y Almond R.G. (1997); Graphical Models and Computerized Adaptive Testing. CSE Technical Report 434.

Mislevy, R. y Gitomer, D.H (1996): The Role Of Probability Based Inference in an Intelligent Tutoring System. CSE Technical Report 413.

Nebel, E. and Masinter, L (1995): RFC 1867 : Form based File Upload in HTML. http://sunsite.auc.dk/RFC/rfc/rfc1867.html

Owen R. J. (1975): A Bayesian Sequential Procedure for Quantal Response in the Context of Adaptive Mental Testing. Journal of the American Statistical Association.

Pescador, F. y Arriaga, J. (1996): Sistema de Autor Orientado Al Refuerzo y Evaluación. En Díaz, A. y fernandez, I. (Eds.): Computer Aided Learning and Instruction in Science and Engineering.

PHP/FI Home Page. URL: http://www.iquest.net/.

Polson, M.C y Richardson J. (1998): Student Model in Foundations of Inteligent Tutoring Systems.

PostgreSQL Home Page. URL: http://www.postgresql.org/

Wainer, H. y Kiely, G. L. (1987): Item clusters and computerized adaptive testing: A case for testlets. Journal of Educational Measurement 24, 189-205.

Warkentyne, H.M.K., Smith, I. Y Forte, E. (1996): Implementation and Evaluation of a WWW Multiple Choice Question Server.

Weber, G. y Specht, M. (1997): User Modeling and Adaptive Navigation Suport in WWW-based Tutoring Systems.

Weiss D. J. y Kingsbury G (1984). Journal of Educational Measurement. Vol 21, Na 4, pp.361-375.

Weiss D. J. y Kingsbury G (1979). An Adaptive Testing Strategy for Mastery Decision. Journal of Educational Measurement. Vol 21, Na 4, pp.361-375.

Welch, R. and Frick, T. W. (1993): Computerized adaptive testing in instructional settings. ETR&D 41, 47-62.