



DEPARTAMENTO DE LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE MÁLAGA

Tesis Doctoral

**Un modelo de evaluación cognitiva basado en
Tests Adaptativos para el diagnóstico en
Sistemas Tutores Inteligentes**

Presentada por:
D. Eduardo Guzmán De los Riscos,
para optar al grado de
Doctor por la Universidad de Málaga.

Dirigida por el doctor:
D. Ricardo Conejo Muñoz,
Profesor Titular de Universidad del
Área de Lenguajes y Sistemas Informáticos.

Málaga, octubre de 2005



DEPARTAMENTO DE LENGUAJES Y CIENCIAS DE LA COMPUTACIÓN
UNIVERSIDAD DE MÁLAGA

El Dr. D. Ricardo Conejo Muñoz, Profesor Titular de Universidad del Área de Lenguajes y Sistemas Informáticos de la E.T.S. de Ingeniería Informática de la Universidad de Málaga,

CERTIFICA:

Que D. Eduardo Guzmán De los Riscos, Ingeniero en Informática, ha realizado en el Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, bajo su dirección, el trabajo de investigación correspondiente a su tesis doctoral titulada:

Un modelo de evaluación cognitiva basado en Tests Adaptativos para el diagnóstico en Sistemas Tutores Inteligentes

Revisado el presente trabajo, estima que puede ser presentado al tribunal que ha de juzgarlo, y autoriza la presentación de esta tesis doctoral en la Universidad de Málaga.

Málaga, octubre de 2005

Fdo.: Ricardo Conejo Muñoz
Profesor Titular de Universidad del
Área de Lenguajes y Sistemas Informáticos.

Esta tesis está dedicada a Mercedes y a mis padres.

*Vive la vida sin hacer daño a los demás;
ilumina a la gente a tu paso.*

Carpe diem, carpe horam.
Horacio

Índice general

Resumen	xiii
I INTRODUCCIÓN	1
1. Introducción	3
1.1. La evaluación en la enseñanza	3
1.1.1. Tipos de evaluación	4
1.1.2. Los tests en la educación	6
1.1.3. Los tests en los Sistemas Tutores Inteligentes	7
1.2. Objetivos	9
1.3. Organización de la memoria	11
II ANTECEDENTES	13
2. La evaluación mediante tests	15
2.1. Breve introducción histórica	16
2.2. Tipos de tests. Ventajas e inconvenientes del uso de tests	16
2.3. Tests Administrados por Computador	17
2.4. Definiciones	20
2.5. Tipos de ítems	21
2.6. Teorías de los Tests	22
2.7. La Teoría Clásica de los Tests	23
2.8. La Teoría de Respuesta al Ítem	26
2.8.1. Clasificación de los modelos basados en la TRI	27
Clasificación basada en el número de rasgos latentes de los que depende los ítems	27
Clasificación basada en el modelo de la CCI	28
Clasificación basada en el tratamiento de la respuesta	30

Taxonomía de Thissen y Steinberg	30
2.8.2. Modelos politómicos	32
2.8.3. Los modelos de respuesta politómicos basados en la TRI	35
Modelo de respuesta graduada	35
Modelo de respuesta nominal	36
Modelo de crédito parcial generalizado	36
Modelo de respuesta para ítems de opción múltiple de Thissen y Steinberg	37
Modelos basados en el proceso de respuesta de Nedelsky	37
Modelo <i>Full Nedelsky</i>	38
Modelo de Revuelta para ítems de opción múltiple	39
2.8.4. Modelos no paramétricos	39
2.9. Los Tests Adaptativos Informatizados	41
2.9.1. Aplicación de la TRI a los TAI	44
Procedimientos de arranque	44
Estimación del nivel de conocimiento	45
Mecanismos de selección de ítems	46
Criterios de finalización	48
2.9.2. Calibración de la CCI	49
Anclaje y equiparación	50
Métodos de calibración de modelos paramétricos	50
Métodos de calibración de modelos no paramétricos: El suavizado núcleo	54
2.9.3. TAI con modelos politómicos	57
Mecanismos de selección de ítems	57
Mecanismos de estimación del conocimiento	58
Criterios de finalización	58
2.10. Los sistemas de tests comerciales	58
2.10.1. Sistemas no adaptativos	59
Intralearn	59
WebCT	59
QuestionMark	59
TopClass	59
I-assess	60
Webassessor	60
C-Quest	60
2.10.2. Sistemas basados en TAI	60
MicroCAT y FastTEST	60

TerraNova CAT	61
CATGlobal	61
2.10.3. Conclusiones	62
2.11. Discusión y conclusiones generales del capítulo	62
3. El diagnóstico en los Sistemas Tutores Inteligentes	65
3.1. Los Sistemas Tutores Inteligentes	65
3.1.1. Breve evolución histórica	65
3.1.2. ¿Qué es un Sistema Tutor Inteligente?	66
3.1.3. Arquitectura de un STI	67
3.1.4. Los Sistemas Educativos Adaptativos para la Web	69
3.2. El modelado del alumno en los STI	70
3.3. El diagnóstico del alumno en los STI	74
3.4. Técnicas para el diagnóstico del alumno en STI	75
3.4.1. Modelos de evaluación basados en heurísticos	75
ELM-ART	75
DCG	78
ActiveMath	79
TANGOW	80
HyperTutor	81
QUANTI	81
3.4.2. Modelos de evaluación basados en tests adaptativos	83
CBAT-2	83
HEZINET	85
Modelo basado en la Teoría de la Decisión	86
Evaluación con tests adaptativos en INSPIRE	88
PASS	89
SKATE	90
3.4.3. Modelos de evaluación basados en lógica difusa	91
Tests adaptativos basados en conjuntos difusos	91
ALICE	93
Evaluación difusa en INSPIRE	93
Entorno de tutorización personalizada, evaluación mediante tests y diagnóstico	95
3.4.4. Modelos de evaluación basados en redes bayesianas	95
ANDES	96
OLAE y POLA	96
Modelo Granularidad-Bayes de Tests Adaptativos	98

Tests Adaptativos Bayesianos	100
3.4.5. Conclusiones	102
3.5. Discusión y conclusiones generales del capítulo	106
III PLANTEAMIENTO	109
4. Un modelo de respuesta basado en la TRI	111
4.1. Descripción general del modelo	111
4.1.1. Modelo discreto	112
4.1.2. Modelo no paramétrico	113
4.1.3. Modelo politómico	114
4.2. Definiciones formales	114
4.3. Discretización del modelo	115
4.4. Curvas características de respuesta y de opción	116
4.5. Tipos de ítems y cálculo de las CCR	117
4.5.1. Ítems verdadero/falso	117
4.5.2. Ítems de opción múltiple	117
4.5.3. Ítems de respuesta múltiple	118
Ítems con opciones independientes	118
Ítems con opciones dependientes	119
4.5.4. Ítems de ordenación	119
4.5.5. Ítems de relación	120
Ítems de emparejamiento	120
Ítems de asociación	121
4.6. Aproximación cuasipolitómica del modelo de respuesta	122
4.7. Calibración de las curvas características	123
Cálculo de las evaluaciones	125
Conversión de las evaluaciones	126
Aplicación del suavizado	126
Estimación de los niveles de conocimiento	127
Refinamiento iterativo	128
4.8. Discusión y conclusiones	128

5. Un modelo basado en TAI para STI	131
5.1. La arquitectura del modelo	132
5.2. Descripción del modelo	134
5.2.1. El módulo experto	134
El modelo conceptual	135
El banco de ítems	137
Los tests	141
Relación entre el número de niveles de conocimiento del modelo conceptual y del test	144
5.2.2. El modelo del alumno	145
5.3. Funcionamiento del modelo	146
5.3.1. La estimación del conocimiento del alumno	148
Inferencia del nivel estimado	152
Criterios de estimación heurísticos	152
Estimación inicial del conocimiento	153
5.3.2. Criterios para la selección de ítems	154
Método bayesiano de la máxima precisión esperada	154
Método basado en la dificultad	155
Método basado en la entropía	157
Método basado en la máxima información	158
5.3.3. Criterios de finalización del test	159
5.4. Calibración de las curvas características del modelo	161
5.5. Conclusiones	162
IV IMPLEMENTACIÓN	165
6. Implementación del modelo: El sistema SIETTE	167
6.1. Tipos de ítems en SIETTE	169
6.1.1. Ítems básicos	169
Ítems de respuesta corta	169
6.1.2. <i>Siettlets</i> o ítems autocorregidos	172
6.1.3. Los ítems generativos	174
6.1.4. La biblioteca de plantillas	177
Ejercicios de pasatiempos	179
6.1.5. Los ítems temporizados	182
6.1.6. Los ítems externos	182
6.2. La base de conocimientos	184

6.3.	El repositorio de modelos del alumno	184
6.4.	El repositorio de profesores	185
6.5.	El aula virtual	186
6.6.	El editor de tests	189
6.6.1.	Construcción del currículo	189
6.6.2.	Creación de ítems	191
6.6.3.	Definición de tests	192
6.6.4.	Tests temporizados	195
6.6.5.	S-QTI	196
6.7.	Analizador de resultados	197
6.8.	Calibrador de ítems	199
6.9.	La interfaz para las conexiones externas	199
6.9.1.	Integraciones simples	201
6.9.2.	Integraciones acopladas	202
Sistemas integrados siguiendo esta aproximación	203	
6.9.3.	Integraciones mediante servicios Web	204
6.10.	Arquitectura de SIETTE	205
6.11.	Detalles de implementación	207
6.12.	Evolución del sistema	209
6.13.	Conclusiones	210
V	EVALUACIÓN	213
7.	Pruebas y evaluación de la propuesta	215
7.1.	Introducción	215
7.1.1.	Evaluación formativa y sumativa	215
7.2.	El simulador	217
7.2.1.	Generación del currículo	218
7.2.2.	Generación del banco de ítems	218
7.2.3.	Generación de los alumnos simulados	220
7.2.4.	Administración simulada del test	220
7.2.5.	Metodología del análisis	221
7.3.	Estudio sobre los tipos de ítems del modelo	221
7.3.1.	Experimento 1: Comparación según los parámetros de las CCI	222
7.3.2.	Resultados	222
7.3.3.	Experimento 2: Comparación según el criterio de selección de ítems	223
7.3.4.	Resultados	224

7.3.5. Discusión	225
7.4. Comparación entre los mecanismos de selección de ítems	225
7.4.1. Experimento 1: Comparación entre el criterio bayesiano y el basado en la dificultad	225
7.4.2. Resultados	226
7.4.3. Experimento 2: Comparación entre el criterio bayesiano y el basado en la entropía	226
7.4.4. Resultados	227
7.4.5. Experimento 3: Comparación entre los criterio bayesiano, basado en la entropía y basado en la información para ítems con diversas propiedades	227
7.4.6. Resultados	229
7.4.7. Discusión	229
7.5. Comparación con el modelo de respuesta 3PL continuo	233
7.5.1. Eficiencia según el número de ítems	233
Experimento 1: Comparación entre los modelos de respuesta dicotomizado y el clásico 3PL	233
Resultados	234
Experimento 2: Comparación entre el modelo de respuesta (politómico) y el modelo clásico 3PL	235
Resultados	235
Experimento 3: Comparación entre los modelos de respuesta (politómico) con selección basada en la entropía y el clásico 3PL	236
Resultados	236
7.5.2. Eficiencia según el tiempo de cómputo empleado	237
Resultados	237
7.5.3. Discusión	242
7.6. Evaluación simultánea de múltiples conceptos	242
7.6.1. Experimento	243
7.6.2. Resultados	243
7.6.3. Discusión	244
7.7. Estudio del método de calibración de ítems	244
7.7.1. Experimento 1: Estudio del valor adecuado para el parámetro de suavizado y de la bondad del método de calibración	245
7.7.2. Resultados	245
7.7.3. Discusión	246
7.7.4. Experimento 2: Estudio de la influencia de las funciones de suavizado	246
7.7.5. Resultados	254
7.7.6. Discusión	258

7.7.7. Experimento 3: Estudio de los heurísticos para calibración	258
7.7.8. Resultados	258
7.7.9. Discusión	261
7.7.10. Experimento 4: Comparación con la propuesta original de Ramsay .	261
7.7.11. Resultados	261
7.7.12. Discusión	265
7.7.13. Conclusiones del estudio	265
7.8. Viabilidad de SIETTE para la calibración de ítems	266
7.8.1. Resultados	267
7.8.2. Discusión	271
7.9. Evaluación formativa del sistema SIETTE	271
7.9.1. Primer prototipo	273
Experimentos realizados	273
Resultados	274
7.9.2. Segundo prototipo	277
Experimentos realizados	278
Resultados	278
7.9.3. Tercer prototipo	279
Experimentos realizados	280
Resultados	281
7.9.4. Cuarto prototipo	284
Experimentos realizados	285
Resultados	288
7.9.5. Conclusiones	291
7.10. Discusión y conclusiones generales del capítulo	296
VI CONCLUSIONES	299
8. Conclusiones	301
8.1. Aportaciones	302
8.2. Limitaciones	306
8.3. Líneas de investigación abiertas	307
A. Lista de abreviaturas	313
Bibliografía	315

Resumen

Los *Sistemas Tutores Inteligentes* (STI), son herramientas para la enseñanza que aplican técnicas de Inteligencia Artificial para guiar al estudiante durante el proceso de instrucción. Utilizan un *modelo del alumno* que representa lo que éste conoce y lo que no conoce del dominio sobre el que está siendo instruido. Este modelo se actualiza durante el proceso de instrucción y es fundamental para determinar en cada momento qué acción debe llevarse a cabo. La estructura que almacena el estado de conocimiento del alumno es propiamente el *modelo del alumno*, mientras que el proceso de razonamiento que actualiza este modelo se denomina *diagnóstico del alumno*. La importancia de un diagnóstico certero y riguroso es vital para el buen funcionamiento de un STI.

Un test es un instrumento de evaluación diseñado para inferir una medida de las capacidades de los sujetos, a través de sus respuesta a un conjunto de preguntas (o ítems). Dentro del diagnóstico en STI, los tests se han popularizado gracias a las ventajas que supone su utilización. Entre éstas destaca el hecho de que son aplicables, en mayor o menor grado, en casi todos los dominios. Sin embargo, tanto el modo en el que se construyen como la forma en la que se aplican suelen ser, por regla general, incorrectos, o cuando menos poco rigurosos. Como a cualquier otro instrumento de medida, se le deben exigir ciertas características, a través de las cuales se pueda verificar su idoneidad.

Los *Tests Adaptativos Informatizados* (TAI), tienen tras de sí un sólido fundamento teórico que garantiza la validez, fiabilidad y objetividad de las evaluaciones que realizan. A través de un TAI se obtiene como resultado la estimación del nivel de conocimiento del alumno. Requieren un menor número de preguntas por test con respecto a los tests convencionales. Esto es debido a que cada pregunta es seleccionada en función de la estimación que hasta ese momento ha realizado el test del conocimiento del alumno. Los TAI se basan en aplicar un algoritmo de evaluación, el cual se sustenta principalmente en una teoría psicométrica denominada *Teoría de Respuesta al Ítem* (TRI). Según ésta, la respuesta que un examinando da a una pregunta, está relacionada con el nivel de conocimiento que éste posee. Esta relación se cuantifica mediante una o más funciones de densidad denominadas *curvas características*.

A pesar de su idoneidad para el diagnóstico del conocimiento, los TAI presentan un conjunto de inconvenientes. El primero de ellos es que sólo pueden ofrecer como resultado, en cada test, una única valoración del conocimiento del alumno. Además, en aquellos tests en los que se ven involucradas preguntas sobre distintos conceptos, los criterios de selección suelen recurrir a heurísticos, lo que pone en tela de juicio el rigor de los resultados. Otro problema de los TAI es que son muy costosos, en cuanto al tiempo de desarrollo. Antes de que un TAI esté operativo, es necesario calibrar las curvas características de cada pregunta. En general, los procedimientos que existen con este fin son muy costosos, puesto que requieren disponer de una muestra poblacional de examinandos de tamaño considerable.

En el ámbito de los STI, existen diversas propuestas en las que se hace uso de los TAI pero, en general, no aprovechan adecuadamente las características y ventajas que éstos pueden ofrecer al diagnóstico del conocimiento. Los modelos de diagnóstico cognitivo persiguen un objetivo más ambicioso que los TAI, puesto que intentan proporcionar un conjunto de inferencias más sofisticado que la simple estimación numérica del conocimiento del alumno en un único concepto que proporciona un TAI. La solución propuesta en esta tesis se basa en la definición de un modelo de respuesta basado en la TRI y de otro de diagnóstico cognitivo, cuya herramienta de evaluación son los TAI. La propuesta incluye diversas características que permiten solventar los problemas de los TAI con respecto al diagnóstico en los STI. Asimismo, el modelo de respuesta definido proporciona un fundamento teórico al de diagnóstico. Permite combinar en un mismo test diferentes tipos de preguntas, e incluye un algoritmo para la calibración de éstas, que economiza los requisitos iniciales que este tipo de algoritmos necesitan para poder ser aplicados. Tanto el modelo de respuesta como el de diagnóstico han sido implementados en un sistema Web, SIETTE, que es capaz de operar como una herramienta de evaluación independiente, o bien como un módulo de diagnóstico dentro de la arquitectura de un STI, gracias a un conjunto de protocolos que se han implementado con este fin.

Parte I

INTRODUCCIÓN

Capítulo 1

Introducción

*Nuestros planes fracasan cuando no tienen un propósito.
Cuando un hombre no sabe qué es lo que pretende,
ningún viento soplará a su favor.*
Lucio Anneo Séneca

1.1. La evaluación en la enseñanza

Una definición genérica de *evaluación* es la propuesta por Cooley y Lohnes (1976) y citada en (Mark y Greer, 1993), según la cual, *la evaluación es un proceso mediante el cual se recopila y se transforman datos relevantes en información para la toma de decisiones*. Asimismo, Winne (1993) la define como un esfuerzo sistemático de recopilar e interpretar, basándose en algún principio, información con la cual valorar la valía de un determinado procedimiento.

Más concretamente, la *evaluación educativa* es un proceso en el que se realizan inferencias sobre lo que el alumno sabe, basándose en evidencias derivadas de observaciones sobre lo que éste dice o hace en determinadas situaciones (Pellegrino et al., 2001). Representa una parte fundamental del aprendizaje, ya que ofrece una medida de lo que ya ha sido aprendido (Patel et al., 1998), así como un medio eficaz para identificar lagunas en el conocimiento de los estudiantes.

La evaluación educativa ha nacido y se ha desarrollado, en el siglo XX, al amparo de la Psicología Experimental (Ruiz de Pinto, 2002). Según la Real Academia de la Lengua Española, el término *evaluación* (desde el punto de vista educativo) se define como la acción de *evaluar*; donde a su vez, *evaluar* se define como: *estimar los conocimientos, aptitudes y rendimiento de los alumnos*.

Según Ruiz de Pinto (2002), la evaluación puede considerarse como una acción racional dotada de propósito. Se concibe, por tanto, como una actividad sistemática indispensable que se integra de lleno en el proceso de enseñanza y aprendizaje; y que a su vez es previa a toda acción contundente a elevar el nivel de la educación (De la Garza Vizcaya, 2004); es decir, su finalidad es la optimización del proceso educativo. Por consiguiente, el fin de la evaluación es la emisión de juicios, que antecederán a las decisiones y a la acción humana (Cronbach, 1963). Otra característica importante, es que es esencialmente comparativa

y, por tanto, supone la adopción de un conjunto de estándares y la especificación de una muestra de referencia, frente a la cual se compara al individuo evaluado.

En la evaluación, utilizando determinados criterios, se obtiene información acerca de un fenómeno, situación, objeto o persona; se emite un juicio sobre lo observado; y se adoptan una serie de decisiones referentes al mismo. De esta forma, desde el punto de vista educativo, su objetivo principal es proporcionar la máxima información, para mejorar el proceso educativo, reajustando los objetivos, revisando críticamente los planes, programas, métodos y recursos, facilitando la máxima ayuda y orientación a los alumnos.

Según Escudero (2003), desde la antigüedad se han creado y usado procedimientos instructivos en los que los profesores utilizaban referentes para evaluar, diferenciar y seleccionar a estudiantes, sin una teoría explícita. Buenos ejemplos del uso de mecanismos de valoración de individuos en la antigüedad, son los empleados en la China imperial de hace tres mil años (2200 a.C.), donde se utilizaban métodos de evaluación para seleccionar a funcionarios; o el que es quizás el tratado de evaluación más importante de la antigüedad, el *Tetrábiblos*, atribuido a Ptolomeo (siglo II). Posteriormente, Cicerón y San Agustín introducen en sus obras conceptos y planteamientos evaluadores. A partir del siglo XVIII, y coincidiendo con un incremento en la demanda de acceso a la educación, se pone de manifiesto la necesidad de comprobar los méritos individuales. Por este motivo, las instituciones educativas empiezan a elaborar e introducir normas sobre la utilización de exámenes. Es principalmente en el siglo XIX cuando aparece la evaluación como actividad y técnica con el nombre de *examen* (Ruiz de Pinto, 2002). El examen pretendía valorar los conocimientos que poseían los alumnos después de la enseñanza impartida. Constituía por tanto, un valioso instrumento didáctico para controlar el aprendizaje de los estudiantes, a la vez que un medio de información de la manera en la que se desarrolló la actividad académica, para revisarla y reorientarla.

Posteriormente, a comienzos del siglo XX aparece el término *test* reemplazando al de examen. El test se considera entonces, como un instrumento científico válido y objetivo, que podría determinar gran cantidad de factores psicológicos de un individuo, tales como la inteligencia, las aptitudes e intereses, el aprendizaje, etc.

1.1.1. Tipos de evaluación

Cuando se llevan a cabo evaluaciones, es sumamente importante tener muy claro cuáles son los objetivos finales que se pretenden. Éstos determinarán el tipo de informaciones que se consideran pertinentes para evaluar, los criterios que se tomarán como punto de referencia, los instrumentos utilizados y la ubicación temporal de la propia actividad evaluadora (Molnar, 2005).

En McCormack y Jones (1997) se esbozan, a grandes rasgos, las líneas que deben seguirse: determinar el conocimiento del alumno antes o durante una lección, estimularle para que contemple el material que ha estudiado, ofrecerle la posibilidad de revisar los conceptos aprendidos, facilitar la posibilidad de que pueda indicar si los ha entendido con suficiente claridad, etc.

Scriven (1967) parece ser el primero en distinguir entre evaluación formativa y evaluación sumativa¹. Posteriormente, Bloom et al. (1971) agregan una nueva categoría, la denominada

¹Scriven hace esta distinción aplicada a la evaluación de sistemas (lo que en inglés se denomina *evaluation*). Aún así, esta clasificación es utilizada en la actualidad por diversos autores (Angelo y Cross, 1993; Black y William, 1998; Bardes y Denton, 2001) para referirse a la evaluación sobre el conocimiento

evaluación inicial o diagnóstica. Estos tres tipos de evaluación no son excluyentes entre sí, por el contrario, son complementarios, y cada uno desempeña una función específica en el proceso de enseñanza y aprendizaje. De esta forma, la evaluación puede clasificarse según el propósito con el que se realiza, es decir, que responde a la pregunta de *para qué*; y está relacionada con el *cuándo* se evalúa. Los tres tipos de evaluación, dependiendo de su función, se enumeran a continuación (Molnar, 2005):

- *Evaluación predictiva, inicial o diagnóstica:* Su finalidad es predecir un rendimiento o determinar el nivel de aptitud previo al proceso educativo. Busca determinar cuáles son las características propias del alumno, antes de que comience el proceso de aprendizaje. El objetivo es, por tanto, ubicarlo en su nivel, clasificarlo y adecuar individualmente el nivel de partida del proceso educativo.
- *Evaluación formativa o de asistencia al aprendizaje:* Se lleva a cabo al finalizar cada tarea de aprendizaje. Su objetivo es inferir los logros obtenidos y, eventualmente, advertir dónde y en qué nivel existen dificultades de aprendizaje. Permite a su vez, la búsqueda de nuevas estrategias educativas. Aporta además, una retroalimentación permanente al desarrollo del proceso de instrucción.
- *Evaluación sumativa, recapitulativa o de logros individuales:* Analiza la cantidad y la retentiva del aprendizaje después de haber completado una unidad de instrucción (Kommers et al., 1996). Su finalidad es obtener un balance del aprendizaje realizado por el alumno, tras finalizar un determinado curso. Sus objetivos son calificar en función de un rendimiento, otorgar una certificación, determinar e informar sobre el nivel alcanzado, etc.

Como se puede apreciar, independientemente de la definición de evaluación que se tome, siempre aparecen involucrados dos componentes: el sujeto (u objeto) evaluado y el criterio (o criterios) que se utiliza como referente (Coll, 1997). Un criterio es un *principio al que se hace referencia, y que permite distinguir lo verdadero de lo falso*, o más precisamente *una característica o una propiedad de un sujeto u objeto, de acuerdo a la cual se formula un juicio de apreciación* (de Landsheere, 2004). Por consiguiente, permite comparar una acción o comportamiento, en relación a otra, que enuncia las reglas del primero y autoriza su evaluación (Molnar, 2005).

Los criterios que definen la bondad de los logros obtenidos por un alumno pueden ser cualitativos o cuantitativos. Los primeros suelen expresarse de forma absoluta, es decir, sin grado alguno: todo o nada, el logro es o no es. Los segundos aceptan un escalado o graduación. Cuanto mayor es la complejidad de los procesos cognitivos, afectivos y psicomotrices de la tarea evaluada, más difícil será la definición de los criterios, porque muchas veces éstos están preimpuestos (de Landsheere, 2004).

Según Glaser (1963), en función del criterio utilizado, pueden distinguirse dos tipos de evaluación (Molnar, 2005):

- *Evaluación normativa:* Significa comparar el resultado del alumno con los de una población o grupo al que pertenece. Requiere el establecimiento de una escala de referencia, obtenida a partir de estudios estadísticos de rendimiento, con el objetivo de obtener una calificación. El criterio es *externo*, puesto que se utiliza un baremo más

o las aptitudes de individuos (educativa), lo que en inglés se denomina mediante un sustantivo diferente: *assessment*.

o menos ajeno al sujeto evaluado. Por esta razón, este tipo de evaluación se utiliza para ubicar a los examinandos en escalas, atribuirles un lugar dentro de los grupos, certificar los niveles en función de la norma o el grupo y predecir futuros resultados.

- *Evaluación (con referencia) a un criterio:* Busca la comparación del alumno con sus propios resultados, en las mismas pruebas o en relación a un criterio fijado de antemano. Se valora principalmente el progreso que éste ha realizado, independientemente de escalas, y se mide su progreso hacia el objetivo propuesto y la distancia que lo separa de él. Constituye la base, a partir de la cual se tomarán ciertas decisiones pedagógicas. Por este motivo, este tipo de evaluación, permite aplicar lo que se denomina una *pedagogía por objetivos* donde, tras haber analizado las necesidades y posibilidades potenciales del estudiante, los retos se expresan en términos operativos (*el alumno será capaz de*). A diferencia de la evaluación anterior, ésta es *interna*, en la medida que no es ajena al individuo.

En esta tesis se plantea desarrollar un modelo cognitivo que permita evaluar al alumno de forma predictiva, al comienzo del proceso de instrucción, para determinar su conocimiento inicial. Esta propuesta deberá también permitir la evaluación formativa durante el proceso de instrucción, de forma que sus resultados sirvan de retroalimentación al propio proceso de instrucción. Finalmente, deberá inferir el estado de conocimiento del estudiante tras finalizar el proceso de instrucción, mediante una evaluación sumativa.

El modelo planteado en esta tesis deberá realizar evaluaciones normativas, que permitan inferir el estado de conocimiento del alumno basándose en una muestra de referencia. Asimismo, deberá permitir abordar evaluaciones con referencia a un criterio, que faciliten el estudio de su evolución durante un proceso de instrucción.

1.1.2. Los tests en la educación

El uso de tests nace de la necesidad de adquirir instrumentos de apreciación objetiva sobre diferentes facultades individuales (Planchard, 1986). Según Pila Teleña (1988), la palabra *test* procede del latín *testa*, que quiere decir prueba; de allí su amplia difusión como término que identifica las herramientas y los procedimientos de evaluación. Santisteban (1990) define un test como un instrumento de medición, diseñado para inferir una medida de las capacidades de los sujetos, a través de sus respuesta. Los tests están formados por preguntas (o ítems).

En general, el uso de tests está muy difundido. En nuestros días, los tests de inteligencia, aptitudes, rendimiento, personalidad y otros procedimientos de medida, forman parte del entorno cotidiano de las personas (Martínez Arias, 1995). Los resultados se aplican en decisiones importantes como promociones y clasificaciones escolares, admisión en centros académicos y de trabajo, y elección de actividades y profesiones. De hecho, países como los EE.UU. hacen uso extensivo de este mecanismo de evaluación en su enseñanza reglada.

Un test es, por tanto, una herramienta de evaluación que se caracteriza por tener las siguientes ventajas: tiene carácter genérico, esto es, es aplicable prácticamente a cualquier dominio (siempre que éste sea declarativo); requiere relativamente poco esfuerzo para llevarse a cabo, e igualmente para corregirse, a diferencia de los exámenes, en los que los alumnos habitualmente deben desarrollar un cierto epígrafe; y da menos pie a interpretaciones personales, tanto por parte del examinando como del profesor, que pudieran poner en entredicho los resultados.

Por el contrario, entre las desventajas que presentan los tests, una de ellas es que no son aptos para tratar procedimientos, es decir, no son aplicables a dominios procedimentales, en los que la evaluación no se limita únicamente a comparar el resultado obtenido por el alumno, con el correcto. Para trabajar adecuadamente con este tipo de dominios, es necesario interpretar las fases realizadas por el examinando y el resultado.

Otra desventaja del uso de los tests reside precisamente en el modo en el que éstos suelen aplicarse. Aunque este mecanismo de evaluación es muy utilizado hoy en día, tanto el modo en el que se construyen como la forma en la que se aplican suelen ser, por regla general, incorrectos. A los tests, como a cualquier otro instrumento de medida, se le deben exigir ciertas características, a través de las cuales se pueda verificar su idoneidad científica (Molnar, 2005):

- *Validez*: La validez de un test no está en función de sí mismo, sino de la aplicación que va a realizarse de él. Se dice que un test posee esta propiedad cuando mide lo que realmente se propone. Según Litwin y Fernández (1977), es precisamente el grado en el cual éste valora aquello que quiere medir. Por ello, los procedimientos existentes para determinar su validez, se basan en establecer la relación entre sus resultados y otros hechos observables, que estén en relación directa con el tipo de capacidad que se intenta evaluar.
- *Fiabilidad*: Hace referencia a la precisión de la medida, independientemente de los aspectos que se pretenden medir. Es la capacidad de un test para demostrar estabilidad y consistencia en sus resultados, y cuando cumple esta propiedad, si se aplica dos o más veces a un individuo, en circunstancias similares, obtenemos resultados análogos. Debe tener en cuenta factores externos que pudieran afectarla como, por ejemplo, el estado de ánimo del evaluado.
- *Objetividad*: Una test es objetivo si sus resultados son independientes de la actitud o apreciación personal del observador. Es el grado de uniformidad con que varios individuos pueden administrar un mismo test. La objetividad garantiza la fiabilidad de un test. Pueden definirse tres ámbitos en los que debe aplicarse: (a) *Objetividad de realización*, en la construcción, aplicación, explicación, descripción e instrucciones del test. (b) *Objetividad de evaluación*, puede ser métrica (utilizando un sistema internacional de medidas) o calificadora (cuando la evaluación es subjetiva). (c) *Objetividad de interpretación*, los grados de valoración de una prueba pueden variar desde objetiva, si ésta se realiza en unas condiciones muy precisas, hasta subjetiva, si el margen de interpretación es muy amplio.

El modelo de evaluación propuesto en esta tesis se basa en la administración de tests. Uno de sus objetivos principales es que las inferencias que realice no se basen en apreciaciones personales (carentes de rigor científico), sino que posean una sólida base teórica, y que cumplan las propiedades de validez, fiabilidad y objetividad, anteriormente descritas.

1.1.3. Los tests en los Sistemas Tutores Inteligentes

En los comienzos de la *Inteligencia Artificial* (IA), el objetivo principal era replicar en una máquina el nivel de inteligencia humano (Brooks, 1991). Posteriormente, y ante la magnitud del problema, estos ambiciosos objetivos iniciales se fueron relajando. Actualmente, los trabajos en este campo se centran, principalmente, en el desarrollo de asistentes inteligentes aplicados a tareas humanas. Dentro de este ámbito, se han desarrollado interfaces

de usuario inteligentes que se adaptan, de forma dinámica, a las necesidades del usuario. Un caso particular de este tipo de interfaces es el software adaptativo, en el que el nivel de dificultad de una aplicación se ajusta según el histórico de resultados anteriores del alumno (Jettmar y Nass, 2002).

Anteriormente se ha puesto de manifiesto la indudable necesidad de la evaluación como medio para mejorar la calidad de los procesos de enseñanza y el aprendizaje en general. La necesidad de disponer de mecanismos de evaluación efectivos es esencial dentro de cualquier proceso de instrucción. A través de la evaluación, es posible identificar lo que el alumno sabe, así como sus puntos fuertes y puntos débiles; observar su propio proceso de aprendizaje; y decidir en qué dirección dirigir ese proceso de aprendizaje (Gouli et al., 2001). La evaluación debería permitir adaptar el proceso de instrucción a las características individuales de cada alumno, detectar sus puntos débiles para poder corregirlos y tener un conocimiento completo y preciso de cada uno. Esto, según Molnar (2005), contribuye a democratizar la enseñanza.

Uno de los campos de aplicación de la IA dentro del ámbito de la educación, es precisamente la evaluación. Una de las principales ventajas de un sistema de evaluación inteligente es que puede contribuir a reducir el tiempo y el esfuerzo empleado por un profesor en llevar a cabo la tarea mecánica y repetitiva de evaluar numéricamente a un alumno. Asimismo, el uso de sistemas de evaluación inteligente permite suministrar al examinando un refuerzo, es decir, una guía, inferida por sus resultados en la evaluación, que permita orientar al alumno en su instrucción. Esto equivale, por tanto, a una tutorización individual (Patel et al., 1998).

Dentro de los *Sistemas Tutores Inteligentes* (STI), herramientas de enseñanza que aplican técnicas de IA (principalmente) para guiar al estudiante durante el proceso de instrucción; la evaluación se utiliza como parte del proceso de diagnóstico (o inferencia) del estado de conocimiento del alumno, ya que esta información, estimada a través de la evaluación, puede utilizarse para guiar la adaptación del sistema (Gouli et al., 2001).

Dentro del diagnóstico en STI, los tests se han popularizado por las ventajas que supone su utilización, y que han sido enumeradas en la sección anterior. El principal problema de aplicación de los tests en el ámbito de los STI es que muchos investigadores no tienen en cuenta una de las premisas fundamentales del diagnóstico: debe ser certero, puesto que de sus resultados va a depender la adaptación del proceso de instrucción. Esto lleva a muchos investigadores en STI a utilizar ciertos tests en los cuales utilizan criterios de evaluación heurísticos, y por tanto, carentes de garantía, validez, fiabilidad, etc.

Existen otro tipo de tests, los *Tests Adaptativos Informatizados* (TAI), que tiene tras de sí un sólido fundamento teórico, que garantiza, entre otras cosas, la validez, fiabilidad y objetividad de los resultados de la evaluación. A través de un TAI se obtiene como resultado la estimación del *nivel de conocimiento* (o denominado de forma genérica, *rasgo latente*) del alumno, que es independiente del test utilizado; es decir, independientemente del TAI administrado, el nivel de conocimiento inferido debe ser el mismo (siempre que no medie, entre administraciones del test, proceso de aprendizaje alguno). Además, los TAI requieren un número menor de preguntas con respecto a los tests convencionales. Esto es debido a que, generalmente, en ellos las preguntas se muestran de una en una, y éstas se seleccionan en función de la estimación del conocimiento del examinando en ese momento.

Los TAI se basan en aplicar un algoritmo de evaluación, el cual se sustenta principalmente en una teoría psicométrica denominada *Teoría de Respuesta al Ítem* (TRI), que establece que la respuesta que un individuo da a una pregunta, está relacionada con el nivel de conocimiento que éste posee. Esto se cuantifica mediante una o más funciones de densidad denominadas *curvas características*.

A pesar de su idoneidad para el diagnóstico del conocimiento del alumno, gracias a su sólido fundamento teórico, los TAI presentan un conjunto de inconvenientes. El primero de ellos es que, sólo pueden ofrecer como resultado, en cada test, una única valoración del conocimiento del examinando. Además, en aquéllos en los que se ven involucradas preguntas sobre distintos conceptos, los criterios adaptativos de selección existentes no son capaces de asegurar una selección balanceada que garantice que el nivel de conocimiento inferido, sea realmente representativo. Normalmente, para garantizar el balanceo en contenido se suele recurrir a heurísticos, lo que pone en tela de juicio el rigor de los resultados. Otro problema de los TAI es que son muy costosos, en cuanto al tiempo de desarrollo que requieren. Para que estén operativos, es necesario calibrar las curvas características de cada una de sus preguntas previamente. En general, los procedimientos que existen para ello son largos y tediosos, puesto que requieren disponer de una muestra poblacional de examinandos de tamaño considerable (como mínimo del orden de la centena). A éstos se les administra un test con las preguntas que se desea calibrar, utilizando heurísticos para la evaluación, y a partir de los resultados se aplica un algoritmo de calibración.

En el ámbito de los STI, existen diversas propuestas en las que se hace uso de los TAI pero, como se pondrá de manifiesto en esta tesis, no aprovechan adecuadamente las características y ventajas que pueden ofrecer los TAI al diagnóstico del conocimiento del alumno.

1.2. Objetivos

Como se podrá apreciar a lo largo de este trabajo, la mayoría de los STI existentes utilizan mecanismos de evaluación basados, principalmente, en heurísticos. Dada la importancia que tiene la evaluación del alumno para dotar de un cierto carácter inteligente a un STI, es necesario buscar métodos de evaluación bien fundamentados, que garanticen un diagnóstico fiable.

Este trabajo se centra en el uso de técnicas de IA aplicadas a la evaluación educativa en el ámbito de los STI. El objetivo principal es el desarrollo y construcción de un nuevo modelo de evaluación genérico basado en TAI, que permita su integración como módulo de diagnóstico en STI. Se ha elegido el paradigma de los TAI por dos razones fundamentales: los tests ofrecen mecanismos de evaluación genéricos, en el sentido de que pueden ser aplicados como medio de evaluación del conocimiento en cualquier disciplina; y se basan en procedimientos bien fundamentados, lo que a priori representa una garantía de que el resultado de la evaluación será correcto y útil como guía en la instrucción.

Asimismo, el modelo de diagnóstico deberá paliar las carencias de los TAI en su aplicación al diagnóstico del estudiante. Como consecuencia, deberá tener las siguientes características:

- Permitir en una única sesión de evaluación (en un solo test) inferir el conocimiento del alumno en varios conceptos.
- La evaluación simultánea de múltiples conceptos en un test debe realizarse de forma balanceada en contenido.
- Maximizar la información obtenida a través de las respuestas del examinando para llevar a cabo la inferencia de su nivel de conocimiento.

Los tests adaptativos utilizan, como mecanismo de inferencia, modelos de respuesta basados en la TRI. En la literatura existen multitud de modelos de respuesta basados en esta teoría aplicables a los TAI. En este trabajo, se estudiarán los más significativos dentro de la TRI, haciendo especial hincapié en los polítómicos. Este tipo de modelos son capaces de maximizar la información acerca del conocimiento del examinando obtenida a través de las respuestas da éste a las diferentes preguntas que forma parte de un test.

Un segundo objetivo de esta tesis es construir un nuevo modelo de respuesta basado en la TRI, que sirva de motor de inferencia al modelo de diagnóstico. Entre las características que éste debe poseer pueden enumerarse las siguientes:

- Debe permitir evaluaciones a diferentes niveles de granularidad, es decir, realizar evaluaciones en diferentes escalas de precisión, en función de las necesidades requeridas por el STI.
- Debe generar el diagnóstico del alumno en un formato fácilmente intercambiable e interpretable por un STI.
- Debe poseer un mecanismo de calibración de ítems que haga que el modelo sea factible. Éste no debe basarse en heurísticos, como algunas propuestas que existen en el ámbito de los STI. Asimismo, debe intentar economizar sus requisitos iniciales, aspecto que, como se pondrá de manifiesto, es una de las principales desventajas de los mecanismos que con este fin se emplean en la actualidad.

Para conseguir estos objetivos, se construirá un nuevo modelo de respuesta basado en la TRI polítómica, implementado de forma discreta con un número variable de posibles niveles de conocimiento. Este modelo será no paramétrico, por lo que las curvas características de los ítems no seguirán ninguna función predefinida, sino que su forma se basará únicamente en las evidencias utilizadas durante la calibración. Será además heterogéneo, permitiendo combinar en un mismo test diferentes tipos de ítems.

Junto con el modelo de respuesta, se propone un algoritmo de calibración inspirado en una propuesta existente, que mejora los resultados que se obtenían con esta última, y que permite calibrar las curvas características de todos los ítems que se definirán como parte del modelo de respuestas. En capítulos posteriores, se procederá a evaluar la bondad de este nuevo algoritmo de calibración, y se llevará a cabo un estudio comparativo del nuevo algoritmo con la versión anterior.

Sobre el modelo de respuesta, se implementará un modelo de evaluación cognitiva para el diagnóstico en STI que, mediante el uso de TAI, permitirá evaluar al alumno tanto en tests sobre un único concepto, como en tests multiconceptuales. Asimismo, y como efecto colateral, mediante un mismo test, se podrá inferir el conocimiento del examinando en otros conceptos relacionados con aquéllos evaluados en el test.

Otro de los objetivos del modelo presentado en esta tesis es que pueda utilizarse como herramienta de diagnóstico del conocimiento del alumno dentro de STI. Por este motivo, éste será implementado en un sistema de evaluación a través de la Web, el sistema SIET-TE, que permita construir y administrar TAI alumnos reales. La idea es que este sistema pueda funcionar de forma independiente o que se integre en otros sistemas como módulo de diagnóstico.

Por último, en capítulos posteriores, se evaluarán el modelo de diagnóstico propuesto en esta tesis, con el objetivo de verificar la bondad de sus resultados, estudiar su comportamiento y rendimiento, y verificar empíricamente sus propiedades. También se realizará un

estudio sobre la herramienta desarrollada con el objetivo de valorar su idoneidad y aplicabilidad a la enseñanza oficial reglada.

1.3. Organización de la memoria

La memoria se compone de ocho capítulos, estructurados en cinco partes. En la primera de ellas, la *Introducción*, compuesta únicamente por este capítulo, se ha realizado una introducción al concepto de evaluación educativa, analizando los diferentes tipos en función de varios criterios. Además, se ha introducido el concepto de test, y se han estudiado las características que éste debe poseer, para poder ser considerado como una herramienta de evaluación con rigor. Finalmente, se ha llevado a cabo una breve aproximación al problema del diagnóstico en los STI, y a cómo los TAI podrían aplicarse como herramienta de evaluación en STI. Asimismo, se ha planteado la problemática del uso de TAI en los STI.

A continuación, en la parte de *Antecedentes*, en el primero de los dos capítulos que la componen, se estudiarán los tests como herramientas de evaluación, haciendo especial hincapié en las ventajas que aportan su administración computerizada. El estudio de los tests se realizará desde el punto de vista formal, es decir, se analizarán las denominadas *teorías de los tests*. En este sentido, se hará especial énfasis, por su relación directa con la propuesta de esta tesis, en la teoría de los TAI. Posteriormente, se estudiará la TRI, mostrando de dónde surge, y se detallará para qué se emplea en los TAI. Se describirán también, los tipos de modelos de respuesta basados en la TRI que existen. A continuación, se realizará un análisis de los TAI basados en modelos TRI polítómicos que, a pesar de tener a priori ciertas desventajas, suponen una maximización de la información sobre los alumnos obtenida a partir de las interacciones con éstos. Finalmente, se estudiarán algunos de los sistemas existentes, en la actualidad, para la elicitation y administración de tests.

En el segundo de los capítulos de la parte *Antecedentes*, se analizará, de forma detallada, qué es un STI, proporcionando una descripción de su arquitectura. Asimismo, se pondrá de manifiesto el problema del modelado y diagnóstico del alumno en los STI. Se realizará un estudio sobre las técnicas de modelado y diagnóstico existentes en el campo de los STI, haciendo especial énfasis en aquéllas aplicadas a dominios declarativos. En este sentido, se describirán algunas de las aplicaciones y propuestas que han hecho diversos investigadores en este ámbito.

La tercera parte es el *Planteamiento*, en la que se describen, en dos capítulos, las dos partes principales de la propuesta realizada en esta tesis. En el primero de ellos, se propondrá un nuevo modelo de respuesta basado en la TRI para ítems polítómicos, que permite maximizar la información suministrada por la respuesta de un alumno a una pregunta. El objetivo es que las estimaciones resultantes sean más precisas y requieran un periodo de tiempo más corto para su resolución. Asimismo, esta propuesta permitirá su aplicación a un conjunto heterogéneo de ítems, que pueden ser combinados dentro de un mismo test. Este modelo incluye un mecanismo propio de calibración inspirado en una técnica anterior, cuyos requisitos son sustancialmente menores que los métodos de calibración más populares.

En el segundo de los capítulos integrantes de la parte de *Planteamiento*, la propuesta anterior se utilizará como modelo de respuesta dentro de un marco genérico para la evaluación mediante TAI en STI. De esta forma, se propondrá un modelo de evaluación cognitiva basado en TAI para el diagnóstico en STI, con modelos del dominio declarativos. Inicialmente, se describirán los componentes del modelo de diagnóstico. Posteriormente, se

procederá a describir cómo lleva a cabo el diagnóstico del alumno. En este sentido, se describirán los mecanismos que se han incluido para la selección de preguntas, para inferencia del conocimiento del examinando y para determinar cuándo debe finalizar el test.

En la tercera parte, la *Implementación*, constituida por un único capítulo, se describe el sistema SIETTE, como la implementación de los dos modelos propuestos. Es una herramienta de evaluación que utiliza como plataforma de difusión Internet. Permite construir y administrar TAI, y además tests convencionales, para poder facilitar así la calibración de los ítems. Esta herramienta de evaluación puede funcionar de forma independiente o integrarse como módulo de diagnóstico en STI con modelos del dominio declarativos. Con este fin, implementa varios protocolos que permite su integración con diversos grados de acoplamiento en otros sistemas.

En la cuarta parte, la *Evaluación*, se estudiarán, de forma empírica, las características de los modelos de respuesta y de diagnóstico propuestos. Para ello, se hará uso de un entorno de generación de alumnos simulados, a los cuales se someterá a tests con diversas características. Esta herramienta de simulación utiliza como base el núcleo central de SIETTE. Los resultados obtenidos permitirán valorar, a través de evaluaciones sumativas, los pros y los contras de los modelos propuestos. Además, dentro de este mismo capítulo, se llevará a cabo una evaluación formativa del propio sistema SIETTE.

Por último, en la parte de *Conclusiones*, se intentarán desgranar las contribuciones que supone este trabajo, así como sus posibles limitaciones. Igualmente, se propondrán las diversas líneas de investigación que se abren como consecuencia de este trabajo.

Parte II

ANTECEDENTES

Capítulo 2

La evaluación mediante tests

*Saber que se sabe lo que se sabe
y que no se sabe lo que no se sabe;
he aquí el verdadero saber.*
Confucio

Un mecanismo de evaluación muy extendido por su generalidad, facilidad de implementación y corrección automática es el uso de tests. Un test, según la definición de la Real Academia de la Lengua Española, es *una prueba destinada a evaluar conocimientos o aptitudes, en la cual hay que elegir la respuesta correcta entre varias opciones previamente fijadas*. En su forma más simple, se compone de un conjunto de instrumentos de medida (preguntas o tareas), que un cierto sujeto debe completar, y que suelen recibir el nombre genérico de *ítems*. Desafortunadamente, la palabra "test" suele asociarse con un examen el cual se aprueba o se suspende. En el contexto de la psicología, este término va más allá. Los tests son considerados herramientas de descripción, en vez de meros instrumentos para juzgar.

Hasta la aparición de los primeros Computadores Personales (PC), e incluso hasta hoy en día, los tests se realizaban sobre papel. Son los denominados *tests de papel y lápiz* (en inglés, *paper-and-pencil tests*). Se componen de una colección de preguntas, en general igual para todos los alumnos, en los que tienen que leer un cierto enunciado y marcar aquellas respuestas que consideran correctas.

Como se verá en el capítulo siguiente, en el ámbito de los STI, el objetivo principal de los tests, es medir lo que se denomina *nivel de conocimiento* del alumno. Éste puede definirse como una medida cuantitativa del estado de conocimiento del alumno acerca de un determinado concepto o materia. Se suele medir mediante una escala numérica con un rango predefinido. De tal forma que, valores pequeños representarán que el examinando posee carencias en el conocimiento de la materia, mientras que valor altos expresan que la domina.

Desde el punto de vista psicológico, se han elaborado teorías matemáticas o estadísticas para medir los rasgos psicológicos que permiten inferir el nivel del examinando en éstos a partir de su rendimiento observado. La rama de la psicología que se encarga del estudio de estas teorías se denomina *psicometría*, que es, por tanto, la ciencia que se ocupa del estudio de aspectos psicológicos como personalidad, habilidad, aptitud o conocimiento.

Según Muñiz (2002), puede verse como el conjunto de métodos, técnicas y teorías implicadas en la medición de variables psicológicas.

2.1. Breve introducción histórica

Los primeros intentos de medir las diferencias entre las características psicológicas de los individuos, pueden encontrarse alrededor del 400 a.C. cuando Hipócrates intentó definir cuatro tipos básicos de temperamento, explicados según el predominio de un determinado tipo de fluido corporal. El primer intento de medir científicamente las diferencias entre las habilidades mentales de los individuos fue realizado por Sir Francis Galton en el siglo XIX, quien intentó mostrar que la mente humana podía clasificarse de forma sistemática en diferentes dimensiones.

En cuanto a los primeros tests, según Escudero (2003), éstos datan de 1845. En esa fecha, en los Estados Unidos, Horace Mann comienza a utilizar tests escritos como técnica de evaluación, aunque no se trata de una evaluación sustentada en un enfoque teórico. En 1879, Rice publica lo que se suele señalar como la primera investigación evaluativa en educación. Se trataba de un análisis comparativo en escuelas norteamericanas, sobre el valor de la instrucción en el estudio de la ortografía, utilizando como criterio de medida las puntuaciones obtenidas en los tests. A principios del siglo XX comenzaron a utilizarse tests como mecanismo de evaluación (van der Linden y Pashley, 2001), y a lo largo de todo el siglo, su uso siguió una progresión de refinamiento y estandarización. En los Estados Unidos, el *College Board* (asociación norteamericana encargada de la elaboración de tests de acceso a la Universidad) utilizó su primer test en 1901. Este organismo, en 1926, administró su primer SAT (*Scholastic Aptitude Test*), test estándar que evalúa si un alumno es apto para ingresar en la Universidad. A partir de entonces, y hasta nuestros días, además de SAT, multitud de tests han sido estandarizados.

2.2. Tipos de tests. Ventajas e inconvenientes del uso de tests

Desde el punto de vista psicométrico, los tests pueden tener diversos usos. Por una parte, los *tests de habilidad* (en inglés, *ability tests*) miden el potencial de una persona, como por ejemplo, la pericia necesaria para conseguir un trabajo. Este tipo de pruebas de evaluación no deben confundirse con los denominados *tests de logros* (en inglés, *attainment tests*), cuyo objetivo es específicamente medir lo que un individuo ha aprendido. La diferencia entre estos dos tipos de tests radica en que los últimos son retrospectivos, esto es, se centran en lo que el individuo ha aprendido; mientras que los primeros son prospectivos, se basan en determinar lo que la persona es capaz de hacer en un futuro.

Otro tipo de tests son los *de aptitud*. No existe una forma clara de distinguir entre éstos y los de habilidad. La diferencia puede establecerse considerando la habilidad como una aptitud subyacente, y la aptitud como algo más relacionado con un trabajo que la habilidad. Por último, los *tests de personalidad*, como su nombre indica miden las características individuales de cada persona. La personalidad puede definirse como *los aspectos estables que perduran en un individuo, y que permiten distinguirlo del resto de personas, haciéndolo único, pero que a su vez permiten compararlos con ellos*.

Como se puede apreciar, la diferencia entre tests de logros, de habilidad y de aptitudes, en ocasiones, difícilmente identificable. Este trabajo se centra en los tests aplicados al ámbito de la educación. El objetivo será medir el nivel de conocimiento del alumno en una o más materias, sin en principio cuestionarse si ese conocimiento es algo subyacente (test de aptitud); si es algo que ha adquirido únicamente gracias a un proceso de aprendizaje llevado a cabo en un tiempo más o menos cercano (test de logros); o si es algo que le va a permitir realizar una actividad futura (test de habilidad). Por este motivo, a lo largo de esta memoria, se utilizarán simplemente el término test como medida instrumental, dejando relegada esta discusión al uso concreto de esta herramienta.

Según Brusilovsky y Miller (1999), dentro de los sistemas educativos actuales, los tests representan uno de los componentes más utilizados para la evaluación. Los tests de evaluación se utilizan en este ámbito como instrumento para medir el conocimiento del alumno en una determinada disciplina. La utilidad de los test es si cabe más importante en los sistemas de enseñanza adaptativos. La evaluación basada en tests es esencial para lograr un proceso de aprendizaje óptimo (Anderson et al., 1989) en aquellos sistemas en los que, junto con la corrección de cada pregunta, se muestra además un refuerzo. Estos últimos son piezas del conocimiento que contribuyen a que el alumno corrija conceptos aprendidos de forma errónea, o bien aprenda otros nuevos, desconocidos hasta ese momento. Son los denominados *tests de autoevaluación*. Además, los resultados de un individuo en un test son una fuente fiable de evidencias sobre si éste ha asimilado los conceptos evaluados. Esta información podrá ser empleada para guiar mejor al estudiante durante el proceso de instrucción en un STI. Mediante tests se pueden actualizar los datos que una herramienta de este tipo posee sobre el estudiante, los cuales se almacenan en el denominado *modelo del alumno*. Por consiguiente, en un STI los tests podrán utilizarse de distintas formas y en diferentes momentos: antes de que dé comienzo el aprendizaje, se hacen uso de los denominados *pretests*, que permiten inferir el conocimiento del alumno antes de que se inicie el proceso de instrucción. Durante éste, pueden utilizarse como complemento adicional mediante los tests de autoevaluación anteriormente mencionados, o como medida de la evolución del aprendizaje del individuo. Finalmente, cuando acaba la instrucción un *post-test* permite medir el grado de aprendizaje que ha sufrido el alumno desde que comenzó.

La principal limitación de los tests en este ámbito es que únicamente son capaces de medir conocimiento declarativo, no siendo útiles como medio de evaluación procedural. Por ejemplo, mediante un test, es posible determinar si el resultado final de un determinado problema es correcto, pero ese test no puede evaluar si el procedimiento que alumno aplicó es correcto. De esta forma, si un examinando comete un error de cálculo no se valora de ninguna forma si aplicó el procedimiento correctamente.

2.3. Tests Administrados por Computador

Desde finales de la década de los ochenta, y debido a la proliferación en el uso de PC, empezaron a realizarse tests sobre soporte electrónico. De esta forma, surgieron los denominados *Tests Administrados por Computador* (TAC). Entre las ventajas de este tipo de tests está la posibilidad que ofrecen al alumno de realizar tests en cualquier momento y en cualquier lugar, además de la disponibilidad inmediata de los resultados. Asimismo permiten la inclusión de nuevos formatos de ítems (Parshall et al., 2000), gracias a la posibilidad de incluir contenidos multimedia tales como imágenes, video, audio, etc...; nuevos tipos de ítems, en los que el alumno debe llevar a cabo una tarea de forma correcta; la posibilidad

de realizar evaluaciones de ítems de respuestas abiertas (Wise, 1999) (o de respuesta corta); etc.

Existen diversos tipos de tests informatizados. El criterio utilizado para distinguirlos se basa en el método empleado para la selección de las preguntas que lo componen, es decir, en si éstos se adaptan al comportamiento del alumno durante la realización del test. En (Patelis, 2000) se realiza una clasificación según este criterio, la cual se ha ampliado, incorporando nuevos tipos, tal y como se muestra en la figura 2.1. Como consecuencia, es posible distinguir las siguientes categorías de tests informatizados:



Figura 2.1: Tipos de tests (adaptado a partir de (Patelis, 2000)).

- *Tests lineales* (en inglés, *linear tests*): Se administran de forma no adaptativa. El término lineal viene a representar la naturaleza secuencial de la realización de los ítems en el test. En estos tests, se administran los mismos ítems para todos los alumnos y su número está predeterminado antes del comienzo. Son la versión computacional de los de lápiz y papel. El objetivo es crear un entorno en el que las propiedades psicométricas de los ítems sean las mismas que cuando éstos se administran en papel. Cada ítem se presenta en el mismo orden y de la misma forma que en un test de papel y lápiz. No existen restricciones con respecto a la posibilidad de revisar ítems anteriormente respondidos. El examinando es libre de cambiar respuestas anteriormente dadas.

La ventaja de este tipo de tests es que son fáciles de implementar. La limitación principal está en el hecho de que a todos los alumnos se les administran los mismos ítems y en el mismo orden. Esto supone un riesgo, ya que los examinandos podrían copiarse entre sí, poniendo en entredicho los resultados de su evaluación.

- *Tests lineales "sobre la marcha"* (en inglés, *linear-on-the-fly tests*): A cada examinando se le administra un test diferente de longitud fija. Los ítems son diferentes para cada alumno, y su selección se lleva a cabo con anterioridad al comienzo del test, conforme a una especificación de contenido y propiedades psicométricas. Es necesario por tanto, disponer de una amplia batería de ítems que permitan construir tests diferentes para múltiples alumnos.

Con respecto a los tests lineales, se mejora la seguridad, puesto que a cada individuo se le suministra un test diferente. Se sigue manteniendo además, la ventaja de que los alumnos, durante la realización del test, pueden modificar las respuestas anteriores en cualquier momento.

- *Testlets*: Un testlet es un conjunto de ítems que se consideran una unidad y que se administran juntos. Se construyen dinámicamente durante la realización del test, a partir de la dificultad de los ítems o de alguna especificación dada a priori por los

expertos; por ejemplo, un conjunto de ítems sobre un enunciado común. En este tipo de tests, los alumnos pueden modificar las respuestas en cualquier momento, durante su realización.

- *Tests con referencia a un criterio* (en inglés, *criterion-referenced tests, proficiency tests,mastery tests,basic skill tests*): Estos tests (Hively, 1974) se utilizan para tomar decisiones precisas sobre la aptitud en una cierta disciplina, objetivo, destreza o competencia; es decir, más que cuantificar el conocimiento del alumno, sirven para decidir si éste es apto. Existen diversos modelos para implementarlos. Su mayor ventaja es la eficiencia, ya que los alumnos son clasificados en función de reglas de decisión simples. Cuando incluyen más de un objetivo, los ítems que cubren cada objetivo se organizan en subtests y el rendimiento de los examinandos es evaluado en cada uno de los objetivos (Martínez Arias, 1995, cap. 21). La mayoría de estos tests se construyen realizando un muestreo aleatorio a partir del conjunto (o banco) de ítems, de aquéllos que pertenecen a un dominio particular, y asumiendo que el rendimiento de los alumnos es el mismo que si se hubieran suministrado todos los ítems de ese dominio del banco. La selección de los ítems se hace de forma aleatoria, para cada examinando, y nunca en función de su nivel de conocimiento.
- *Tests Auto-adaptados Informatizados*(TAAI) (en inglés, *Computerized Self-Adapted Tests*): Fueron propuestos a finales de los años 80 (Rocklin y O'Donnell, 1987), y permiten al alumno elegir el nivel de dificultad de las preguntas que le van a ser suministradas. De esta forma, el banco de ítems se divide en categorías (entre cinco u ocho) ordenadas por dificultad. El proceso de aplicación de un test de este tipo es el siguiente: 1) El examinando elige el nivel de dificultad del primer ítem. 2) Se le administra uno de esa categoría elegido al azar. 3) Tras responderlo, se le suministra un refuerzo sobre el resultado, y se pide que vuelva a elegir la dificultad del siguiente ítem. 4) Se repiten los pasos de 2 a 4 hasta que se han aplicado un número determinado de ítems, o bien, se ha estimado el conocimiento del alumno con la precisión requerida. La forma de inferir la estimación de conocimiento es la misma que emplean los Tests Adaptativos Informatizados.

Los TAAI fueron ideados a partir de un análisis de un test relativamente fácil en el que se descubrió que, los individuos con baja ansiedad rindieron peor que los de ansiedad moderada; justo lo contrario que sucedía en un test relativamente difícil (Wise, 1999). Se llegó por tanto a la conclusión, de que la ansiedad es un factor que influye en el rendimiento de los alumnos. Los TAAI fueron ideados para estudiar hasta qué punto los examinandos eran capaces de elegir la dificultad para la que se optimiza su propio rendimiento.

- *Tests Adaptativos Informatizados* (TAI) (en inglés, *Computerized Adaptive Tests*): En ellos, las preguntas que conforman cada test se van seleccionando en función de la respuesta que el alumno haya dado a la pregunta anterior. Cada ítem viene caracterizada por un conjunto de parámetros. Éstos se calculan mediante un proceso denominado *calibración*, a partir de los resultados de tests realizados con esas mismas preguntas de forma no adaptativa. Cuando estos parámetros están bien determinados, se dice que los ítems están bien calibrados.

Para determinar la siguiente pregunta que debe mostrarse al alumno, así como para determinar el procedimiento de inferencia de su conocimiento y cuándo debe finalizar el test, principalmente se utiliza la denominada *Teoría de Respuesta al Ítem* (TRI) (en inglés, *Item Response Theory*). No obstante existen otras propuestas de TAI que

no hacen uso de la TRI, como por ejemplo los tests ramificados, que serán descritos posteriormente en este capítulo. Algunas otras propuestas de TAI no basados en la TRI, serán analizadas a lo largo del siguiente capítulo.

A grandes rasgos, las ventajas principales de este tipo de tests son las siguientes: 1) La eficiencia, que con ítems bien calibrados se traduce en reducciones en más del 50 por ciento del número de ítems requerido para evaluar de forma precisa a un alumno, en comparación con un test no adaptativo. 2) El amplio rango de medida que proporcionan, ya que la escala utilizada para medir el conocimiento es más precisa que en los restantes tipos de tests. 3) La seguridad, debida en gran parte al hecho de que a cada individuo se le suministra un conjunto y número diferente de ítems.

La desventaja principal de los TAI es técnica, ya que antes de poder utilizarlos es necesario haber administrado sus ítems a un número suficiente grande de alumnos de forma no adaptativa, sin utilizar criterios de evaluación basados en la TRI y en un entorno controlado, para poder calibrarlos. Por este motivo, el uso de TAI suele estar restringido a grandes organizaciones con el soporte necesario para poder llevar a cabo la calibración. Entidades como el *Educational Testing Service* (ETS) son los encargados de la creación y administración de algunos de los TAI más prestigiosos como el GRE (*Graduate Record Examination*) (Mills y Steffen, 2000); los tests del CAT-ASVAB (*Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery*) (Sands et al., 1996), la batería de tests sobre aptitud vocacional para las fuerzas armadas. Entre los test que han sido estandarizados quizás destaca, por su mayor difusión y prestigio internacional, el TOEFL (*Test of English as a Foreign Language*), que se utiliza en 88 países. Se trata de un prueba para evaluar el conocimiento de la lengua inglesa, requisito imprescindible para que extranjeros puedan realizar estudios en las universidades norteamericanas. Además, ciertas empresas están migrando sus tests a TAI, tales como Microsoft con su programa de tests de Certificación Profesional de Microsoft (Hudson, 1999), o Novell, que ya han administrado más de un millón de TAI de su programa de certificación (Novell, 2004).

2.4. Definiciones

Antes de proseguir con esta memoria, y para facilitar al lector una adecuada comprensión del texto, se procederá a definir brevemente algunos términos relacionados con la evaluación mediante tests, que serán utilizados con frecuencia a lo largo de este trabajo:

- **Alumno:** También llamado a lo largo de esta memoria *examinando*, *individuo* o *estudiante*. Representa a un sujeto que ha sido, está siendo o va a ser sometido a una evaluación.
- **Test:** Se define como la especificación de una prueba o evaluación. Esta especificación consta de: (i) un conjunto de ítems; (ii) criterios para seleccionarlos, es decir, cómo se eligen aquéllos que van a ser mostrados a cada alumno; (iii) criterios de evaluación, es decir, cómo se infiere el estado de conocimiento del alumno a partir de sus respuestas a los ítems; y (iv) criterios de finalización del test, es decir, cuándo debe finalizar.
- **Ítem:** Osterlind (1990) define ítem de la siguiente forma: *Un ítem de un test (...) es una unidad de medida con un estímulo que genera una respuesta en un formato; y cuyo objetivo es obtener una respuesta de un alumno, cuyo rendimiento en algún aspecto*

psicológico (tal como su conocimiento, habilidad, predisposición, rasgo, etc.) quiere ser inferido. De esta forma, el concepto de ítem incluye no sólo las clásicas preguntas con un enunciado y un conjunto de posibles respuestas, sino también cualquier tipo de ejercicio que permita la evaluación de un individuo en la escala definida en un test.

- **Sesión:** Es en sí la realización de un test por parte de un alumno. A partir de ella puede extraerse la siguiente información: los ítems que han sido administrados, la respuesta que el alumno ha dado, y la calificación que ha obtenido.
- **Banco de ítems:** Colección de ítems a partir de la cual se construyen los tests. Deben tener unas características específicas, las cuales se pondrán de manifiesto a lo largo de esta tesis.

2.5. Tipos de ítems

Dado que esta tesis se centra en la evaluación mediante tests informatizados, en este apartado se van a describir gran parte de los ítems que pueden aparecer en un sistema de tests de este tipo. Estos elementos pueden formar parte de lo que podría considerarse el *modelo de tareas* (en inglés, *task model*) de una herramienta de evaluación.

Existen diversas taxonomías para denominar las diferentes categorías de ítems, aunque ninguna de ellas es suficientemente exhaustiva. En este apartado se realizará una clasificación propia basada en la combinación de las propuestas de Osterlind (1998) y Parshall et al. (2000).

- *Ítems de verdadero/falso* (en inglés, *true/false items*): Son aquéllos en los que los alumnos deben responder en función de la veracidad de cierto enunciado. Las respuestas posibles son sólo dos: verdadero/sí o falso/no.
- *Ítems de opción múltiple* (en inglés, *multiple choice items*): También llamados *ítems de respuesta simple*. Son aquéllos en los que los alumnos tienen que seleccionar una sola de entre un conjunto de posibles respuestas, pudiendo dejar la pregunta en blanco, es decir, sin señalar ninguna respuesta.
- *Ítems de respuesta múltiple* (en inglés, *multiple response items*): Son similares en formato a los anteriores, pero en este caso los examinados pueden seleccionar más de una respuesta de entre el conjunto de alternativas. El número de respuestas que los alumnos deben elegir puede especificarse o no.
- *Ítems de respuesta sobre figura* (en inglés, *figural response items*): Son aquéllos en los que los alumnos tienen que seleccionar una parte de una figura o de un gráfico.
- *Ítems de ordenación* (en inglés, *ordered response items*): Presentan al examinando una lista de elementos que deben ser ordenados de acuerdo a cierto criterio.
- *Ítems de correspondencia* (en inglés, *matching items*): Son ítems en los que a los alumnos se les presentan dos columnas de elementos que deben ser relacionados entre sí dos a dos en función de cierto criterio.
- *Ítems de respuesta corta o abierta* (en inglés, *open-answer, short-answer or fill-in-the-blanks items*): Son aquéllos en los que los examinados tienen que dar una respuesta escrita, relativamente corta, dado cierto enunciado. A excepción del este último no se proporciona ninguna otra información sobre la respuesta.

- *Ítems de completar* (en inglés, *completion items*): Son similares a los anteriores, pero en este caso, los alumnos deben llenar, con respuestas escritas, diversos huecos que han sido dejados en el enunciado.
- *Ítems de redacción* (en inglés, *essay items*): Estos ítems son como los de respuesta corta, pero con la diferencia de que la respuesta puede tener una extensión considerable. Suelen corregirse automáticamente mediante técnicas de procesamiento del lenguaje natural o manualmente por el profesor.
- *Ítems por partes* (en inglés, *multi-part items*): Para la resolución de este tipo de ítems, el alumno debe realizar un conjunto de pasos. El ítem se considera completamente correcto cuando se han realizado todas los pasos correctamente.

2.6. Teorías de los Tests

Cuando se llevan a cabo mediciones de una característica a través de tests, si se intenta medir lo mismo dos veces consecutivas, con frecuencia se obtendrán valores diferentes. Si el instrumento de medida es las dos veces el mismo, esta diferencia puede deberse a un cambio en la propiedad medida. En ocasiones, estas variaciones en el atributo valorado se deben al mero hecho de medir. Hay casos en los que el sujeto de la prueba aprende durante la realización de la misma, y como resultado, al aplicarle una segunda, la medida de sus características difiere. Sin embargo, cuando no se produce realmente un cambio en el atributo, la diferencia entre dos mediciones realizadas puede deberse al denominado *error de medida*.

Las *Teorías de los Tests* (en inglés, *Test Theories*) o *Teorías de la Medida* (en inglés, *Measurement Theories*) surgen a partir de lo que en estadística se conoce con el nombre de *Teoría del Muestreo* (Gonvindarajulu, 1999). Ésta se basa en tomar muestras de una determinada población de elementos, y a partir de ellas se tendrán en cuenta ciertos criterios de decisión. Gracias al muestreo es posible analizar diversas situaciones sobre un ámbito en concreto, como por ejemplo la sociedad.

Las teorías de los tests pueden verse como un conjunto de definiciones, axiomas y suposiciones que permiten, cuando las éstas últimas se cumplen, la estimación de propiedades psicométricas. En este sentido, proporcionan modelos para las puntuaciones de los tests. Éstos permiten obtener, por una parte, la estimación del nivel que poseen los individuos en la característica (o características) medidas por el test; y por otro lado, la estimación de los parámetros de los ítems. El problema central en las teorías de los tests es la relación que existe entre el nivel del individuo, cuyo valor no es observable en el test, y su puntuación observada en el mismo.

Por consiguiente, el objetivo de cualquier teoría de tests es inferir el nivel de un sujeto en la característica o rasgo medido, a partir de la respuestas proporcionadas a los elementos del test. Como consecuencia, hay que establecer una relación entre las características latentes de los individuos y su actuación. Ésta vendrá, en general, descrita por una función matemática. La principal utilidad de la *Teoría de la Medida* es saber hasta qué punto es idónea la estimación realizada del nivel del sujeto. Las distintas teorías difieren precisamente en la función que utilizan para relacionar la actuación observable en el test con el nivel en la característica medida por éste. Esta función permite calcular la característica (o características) que mide el test, y el error que se produce inevitablemente en cualquier proceso de medición.

Según Gulliksen (1961), y tal y como se cita en (Mislevy, 1993), el principal problema de la teoría de los tests es la relación entre la variable que se desea medir a través del test y la puntuación observada tras realizarlo. Este problema es análogo al mito de la caverna de Platón, donde los psicólogos asumen el rol de moradores de la caverna. Ellos sólo pueden conocer la variable medida en el test a través de las sombras (las puntuaciones observadas) proyectadas en el muro de la caverna. El problema reside en cómo hacer un uso efectivo de esas sombras para determinar la naturaleza de la realidad (variable medida en el test).

Las teorías de los tests pueden estar basadas en tests o en ítems. Las dos más importantes se enumeran a continuación:

- *Teoría Clásica de los Tests*: Se fundamenta en la suposición de que la puntuación observada de un alumno se compone del valor real obtenido más la medida del error, las cuales no están correlacionadas. Se tratará con más detalle en la siguiente sección.
- *Teoría de Respuesta al Ítem*: A diferencia de la anterior, se basa en ítems, que se consideran independientes, es decir, la respuesta que un sujeto da a un ítem no está condicionada por la que dio a otros anteriores. Esta teoría se analizará con detenimiento en la sección 2.8.

2.7. La Teoría Clásica de los Tests

La *Teoría Clásica de los Tests* (TCT) (Muñiz, 2002) (también llamada *Teoría de la Puntuación Verdadera*) empezó a utilizarse a comienzos del siglo XX, y lleva aplicándose durante décadas. Inicialmente propuesta por Spearman (1904), más que una única teoría es un compendio de teorías y técnicas *ad hoc* con diversos grados de formalización (Gaviria, 2002).

En la TCT el conocimiento (*habilidad* o *puntuación verdadera*) se define como "el valor esperado obtenido por un alumno en un cierto test". Sea un alumno a que realiza un determinado test i , se expresa de la siguiente forma:

$$Y_{ai} = \tau_{ai} + \varepsilon_{ai} \quad (2.1)$$

donde Y_{ai} es una variable aleatoria que representa la puntuación observada de a al responder a i , también denominada *puntuación en el test*. Se compone de dos partes: la *puntuación verdadera* (τ_{ai}) y el *error de medida* (ε_{ai}), ambas variables latentes no observables. Es decir, Y_{ai} puede calcularse a partir del número de preguntas correctamente respondidas, o mediante otro heurístico. A su vez, ε_{ai} es una variable aleatoria con distribución normal, de media cero y varianza desconocida.

Otra hipótesis de la ecuación 2.1 es que τ_{ai} y ε_{ai} no están correlacionadas. Si se llevan a cabo dos mediciones distintas, los errores de ambas son independientes entre sí: el error que se comete en una medida no guarda ninguna relación con la puntuación verdadera de otra. Como consecuencia, la medida de la habilidad sólo tiene sentido en el marco de un cierto test.

Más formalmente, la TCT hace referencia a un experimento aleatorio de: (1) una persona a de un conjunto de personas Ω_U (la población); y (2) una o más observaciones de un conjunto Ω_O de posibles observaciones. Nótese que, en este caso, las observaciones son de tipo cualitativo: serán las respuestas que a da a las preguntas del test. El conjunto de posibles

salidas de este experimento aleatorio es el conjunto producto $\Omega = \Omega_U \times \Omega_O$. Asimismo, a partir de experimento aleatorio anterior, se puede definir la función U de correspondencia o proyección, $U : \Omega \rightarrow \Omega_U$, que puede considerarse una variable aleatoria cualitativa. Por otra parte, se define también la función Y_i , que permite calcular la puntuación obtenida por un alumno en el test, y que viene descrita de la siguiente forma: $Y_i : \Omega \rightarrow \mathbb{R}$.

Consecuentemente, la variable *puntuación verdadera* τ_{ai} representa los valores condicionales esperados de $E(Y_{ai}|U = a)$ de Y_{ai} , dado el sujeto a . La variable de error ε_{ai} se define a través de la siguiente diferencia: $\varepsilon_{ai} = Y_{ai} - \tau_{ai}$.

A partir de las definiciones formales de las variables de puntuación verdadera y error, pueden derivarse los denominados axiomas de la TCT (Novick, 1966), que son propiedades inherentes a ambas definiciones:

1. *Descomposición de las varianzas*, que es consecuencia directa de independencia entre τ_{ai} y ε_{ai} , y que se formula de la siguiente forma:

$$\sigma^2(Y_{ai}) = \sigma^2(\tau_{ai}) + \sigma^2(\varepsilon_{ai}) \quad (2.2)$$

2. Otras propiedades implícitas en la definición de las variables de puntuación verdadera y error:

$$Cov(\tau_{ai}, \varepsilon_{ai}) = 0 \quad (2.3)$$

$$E(\varepsilon_{ai}) = 0 \quad (2.4)$$

$$E(\varepsilon_{ai}|U) = 0 \quad (2.5)$$

El principal objetivo de la TCT es el estudio de la fiabilidad de la puntuación obtenida en el test. Dados los axiomas anteriores, la fiabilidad se puede definir en función de la varianza, de la siguiente forma:

$$Fiab(Y_{ai}) = \frac{\sigma^2(\tau_{ai})}{Y_{ai}} \quad (2.6)$$

El valor de este coeficiente de fiabilidad oscilará entre cero y uno. Permite comparar diferentes instrumentos de medida aplicados sobre la misma población para determinar su calidad.

A partir de este estudio de la fiabilidad, la TCT responde a las siguientes preguntas: (1) ¿Cómo se puede establecer la correlación entre las dos variables aleatorias una vez que se ha eliminado el error de medida? Es la denominada *corrección por atenuación*. (2) ¿Cuál es el intervalo de confianza de la puntuación verdadera de un sujeto con respecto a la medida considerada? (3) ¿Cuál es el grado de fiabilidad de una medida agregada compuesta por la media (o suma) de varias medidas de la misma característica? (4) ¿Qué grado de fiabilidad puede asumirse en función de la discrepancia entre dos tests diferentes?

Existen diversos modelos basados en la TCT. A continuación se describen brevemente algunos de los más relevantes:

- *Modelo de tests paralelos*: (en inglés, *Parallel Test Model*) Dos tests Y_{ai} y Y_{aj} son paralelos si: (a) Son τ -equivalentes, es decir, si se cumple que: $\tau_{ai} = \tau_{aj}$. (b) Sus variables de error no están correlacionadas. (c) Tienen idénticas varianzas de error.

Como resultado, gracias a la τ -equivalencia, la ecuación 2.1 puede reescribirse de la siguiente forma: $Y_{ai} = \tau + \varepsilon_{ai}$, con lo que el cálculo del error se lleva a cabo de la siguiente forma: $\varepsilon_{ai} = Y_{ai} - E(Y_{ai}|U)$.

Utilizando el modelo de tests paralelos, puede inferirse la fiabilidad de considerar la puntuación obtenida en m tests Y_1, \dots, Y_m como la suma de las puntuaciones en cada uno ellos, esto es, $S = Y_1 + \dots + Y_m$. Ésta se calcula aplicando la denominada *fórmula de Spearman-Brown*:

$$Fiab(S) = Fiab(S|m) = \frac{mFiab(Y_{ai})}{1 + (m - 1)Fiab(Y_{ai})} \quad (2.7)$$

- *Modelo de tests τ -equivalentes esenciales:* Su definición es menos restrictiva que la del modelo de los tests paralelos. Dos tests Y_{ai} y Y_{aj} son *esencialmente τ -equivalentes* si sus puntuaciones verdaderas difieren únicamente en una constante, es decir, $\tau_{ai} = \tau_{aj} + \lambda_{ij}$, $\lambda_{ij} \in \mathbb{R}$. Asimismo, en este modelo se sigue manteniendo la hipótesis de que sus variables de error no están correlacionadas.

De forma análoga al caso anterior, se define un coeficiente de fiabilidad de la puntuación S obtenida tras la administración de m tests. En esta ocasión, es el denominado *coeficiente de Cronbach α* :

$$\alpha = \frac{m}{m - 1} \left(1 - \frac{\sum_{i=1}^m m\sigma^2(Y_{ai})}{\sigma^2(S)} \right) \quad (2.8)$$

Este coeficiente representa el límite inferior de la fiabilidad de S si se asume que los errores no están correlacionados.

- *Teoría de la Generalizabilidad* (en inglés, *Generalizability Theory*): Esta teoría (Shavelson y Webb, 1991) es una extensión de la anterior, en la que se considera que pueden existir múltiples fuentes de errores de medida, denominadas *facetas*. La teoría se basa precisamente en la identificación y cuantificación de estas fuentes de error.
- *Modelo de tests τ -congenéricos:* Se dice que dos tests Y_i y Y_j son τ -congenéricos (Bollen, 1989) si sus puntuaciones verdaderas son combinaciones lineales positivas la una de la otra ($\tau_i = \lambda_{i0} + \lambda_{i1}\tau$, $\lambda_{i0}, \lambda_{i1} \in \mathbb{R}$, $\lambda_{i1} > 0$), y si además sus errores no están correlacionados. Se asume además que $\sigma^2(\tau) = 1$.

En la TCT los ítems vienen caracterizados por dos parámetros: la *dificultad*, que es la proporción de examinandos que responden correctamente al ítem; y el *índice de discriminación*, que se obtiene mediante una correlación entre el ítem y la puntuación total del test.

En el ámbito de la psicología, las medidas son frecuentemente denominadas *reglas de puntuación del test*. Éstas describen cómo se transforman las observaciones en puntuaciones, y son la suma de los puntos obtenidos en los ítems o también pueden calcularse mediante procedimientos más sofisticados. En cualquier caso hay que precisar que la TCT no prescribe cómo hacer estas inferencias. Únicamente descompone este resultado en dos sumandos. Son necesarios estudios adicionales, junto con validaciones empíricas, para decidir si el método de puntuación utilizado para calcular la calificación en un test es significativo. A través de la TCT sólo se pueden extraer las varianzas de su puntuación verdadera y su error.

Esta teoría presenta muchas limitaciones (Hambleton et al., 1991), y aquellos paradigmas que derivan de ella no resultan adecuados para modelar las respuestas a los ítems de un

test. La medida del conocimiento está fuertemente ligada a las características del mismo, y además, no se logra un valor cuantitativo absoluto, sino que lo que se obtiene como resultado depende del test que haya sido aplicado. Esto dificulta enormemente la comparación entre alumnos que hayan hecho tests diferentes (en los que los ítems poseen diversas dificultades). Asimismo, los parámetros de los ítems representan características de una población determinada, no siendo genéricos para cualquier muestra. Por lo tanto, la facilidad o dificultad de un ítem vendrá determinada por los niveles de conocimiento de los individuos considerados, y a su vez, este dato estará condicionado por la facilidad o dificultad del test (de sus ítems). Por último, los supuestos en los que se fundamenta la TCT son bastante difíciles de contrastar empíricamente (Olea y Ponsoda, 2001).

Por el contrario, como ventaja, la TCT es fácilmente aplicable en diversas situaciones, gracias precisamente a la falta de una base teórica robusta (Hambleton y Jones, 1993). Asimismo, otras ventajas citadas por McCallon y Schumacker (2002) son que, en comparación con la Teoría de Respuesta al Ítem y desde el punto de vista de la facilidad para poder ser puesta en práctica, requiere un número menor de examinandos; y que los criterios tradicionales de evaluación utilizados en tests (porcentaje de ítems acertados, puntuación obtenida, ...) cumplen fácilmente sus supuestos.

2.8. La Teoría de Respuesta al Ítem

La *Teoría de Respuesta al Ítem* (TRI) (en inglés, *Item Response Theory*) (Hambleton et al., 1991; van der Linden y Hambleton, 1997; Embretson y Reise, 2000) se encarga de modelar el conjunto de procesos relacionados con la respuesta de un alumno a un ítem (Thissen, 1993). Esta teoría conceptualmente se basa en los primeros trabajos de Thurstone (1925) sobre el concepto de proceso de respuesta y en los trabajos de Lazarsfeld (1950) y Lord (1952) sobre la relación de las variables latentes (no observada) con las respuestas a los ítems y las puntuaciones obtenidas en los tests.

Desde la aparición del libro de Lord y Novick (1968) en el que se introducía la Teoría de la Medida basada en modelos, se ha producido una revolución en el campo de la Teoría de los Test. La TRI se ha convertido en el fundamento principal de la Teoría de la Medida y representa una alternativa a la TCT. Dos hechos principales han favorecido que la TRI se imponga a la TCT como instrumento de medida basado en tests (Gaviria, 2002): su superioridad teórica y conceptual frente a su predecesora, y la existencia de programas de ordenador que facilitan el cálculo de los parámetros de los ítems.

La TRI se apoya en dos principios fundamentales (Hambleton et al., 1991): los resultados obtenidos por un individuo en un test pueden ser explicados mediante un conjunto de factores denominados *rasgos latentes* o *habilidades*, que pueden medirse mediante valores numéricos inicialmente desconocidos. Además, la relación entre los resultados del alumno en el test y sus respuestas a un cierto ítem puede ser descrita mediante una función monótona creciente denominada *Curva (o Función) Característica del Ítem* (CCI).

La CCI es el elemento básico de la TRI, el resto de elementos de esta teoría dependen de ella, y representa la probabilidad condicional de que un examinando, con un cierto rasgo latente estimado (θ), responda correctamente a un ítem. Esta función debe ser conocida para cada ítem del test y, en general, se expresa mediante: $P : (-\infty, \infty) \rightarrow [0, 1]$ en la que el rango de valores fluctúa en la escala en la que se mide θ .

Existen un conjunto de características deseables cuando se aplica la TRI (Hambleton et al., 1991):

- *Invarianza:* Expresa la independencia en las estimaciones. Es decir, que sin tener en cuenta el test que se esté aplicando, el rasgo latente del alumno que está siendo medido, debe ser siempre el mismo. Además, este valor no se modifica durante la realización del test porque se debe asumir que no se produce ningún aprendizaje. Asimismo, en los modelos paramétricos, esta característica es extensible a los parámetros que caracterizan a la CCI. Independientemente del grupo de examinandos que se utilice para llevar a cabo la calibración de cada ítem, se asume que los valores inferidos de sus parámetros deben ser los mismos.
- *Independencia local:* La respuesta a un ítem no debe servir de ninguna forma de ayuda al alumno para responder a otro posterior. Esto se expresa matemáticamente de la siguiente forma:

$$P(U_1, U_2, \dots, U_n | \theta) = P(U_1 | \theta)P(U_2 | \theta) \dots P(U_n | \theta) \quad (2.9)$$

La probabilidad $P(U_1, U_2, \dots, U_n | \theta)$ de que un individuo con un nivel de conocimiento θ responda a un test con un cierto patrón de respuesta U_1, U_2, \dots, U_n , es igual al producto de las probabilidades, $P(U_i | \theta), 1 \leq i \leq n$, de que responda a cada uno de los n ítems dado su nivel de conocimiento.

- *Unidimensionalidad:* Según este principio, el único factor que influye en la respuesta del alumno es el rasgo latente que éste tiene. Esto indica que un modelo basado en la TRI sólo debe medir un único rasgo latente en cada test. Este principio no puede seguirse de forma estricta ya que son muchos los factores que intervienen en la realización de un test y que afectan a su resultado: motivación, ansiedad, habilidades cognitivas, etc. Por consiguiente, esta condición se relaja requiriendo que haya un rasgo dominante sobre el resto. Este factor es la habilidad medida en el test. Aún así, en la actualidad existen modelos que no exigen esta propiedad, en los que se pone de manifiesto la influencia de diversos rasgos latentes en las respuestas de los alumnos a cierto tipo de ítems.

2.8.1. Clasificación de los modelos basados en la TRI

Existen multitud de modelos basados en la TRI. En todos ellos, la probabilidad de una respuesta correcta depende del conocimiento θ del examinando, y de los parámetros que caracterizan al ítem (Martínez Arias, 1995). En este apartado, se clasificarán estos modelos atendiendo a los siguientes criterios: (a) en función del número de rasgos latentes de los que dependen sus ítems de forma simultánea; (b) según el tipo de función que define a las curvas características; y por último, (c) en función de cómo el modelo lleva a cabo el tratamiento de la respuesta proporcionada por el alumno. Es necesario precisar que estos criterios de clasificación no son excluyentes entre sí, es decir, un mismo modelo podrá ser clasificado en más de un grupo.

Clasificación basada en el número de rasgos latentes de los que depende los ítems

Se realiza en función del número de rasgos latentes que un ítem puede evaluar (o proporcionar evidencias) de forma simultánea. Si cada ítem únicamente evalúa un rasgo latente, el modelo se denomina *unidimensional*. En otro caso, se dice que es *multidimensional*. Aunque los primeros son más comunes por su simplicidad, bien es cierto que en los últimos años han

sido desarrollados diversos modelos multidimensionales, gracias a los cuales la respuesta a un ítem puede suministrar información sobre el nivel de conocimiento de un examinando en más de un rasgo latente. En este tipo de modelos, las CCI son multidimensionales, por ejemplo, si un ítem depende de dos rasgos latentes su CCI será bidimensional, si depende de cuatro tetradimensional, etc. Esto, obviamente, complica aún más la ardua tarea de calibrar estas curvas. Entre los modelos multidimensionales más relevantes destacan las propuestas de Segall (1996, 2001), Embretson (1991) y Luecht (1996). Una buena revisión de este tipo de modelos puede encontrarse en (Hontangas et al., 2000), y una comparativa de éstos aplicados a TAI en (Tam, 1992).

Clasificación basada en el modelo de la CCI

El segundo criterio se establece en función de la forma de la CCI. Existen muchas funciones que permiten modelar esta curva. Los primeros que fueron utilizados eran los denominados *modelos normales*, porque estaban basados en la función de distribución normal.

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta-b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \quad (2.10)$$

Éstos presentan la desventaja de que requieren cálculos matemáticos muy costosos. Por este motivo, han sido sustituidos por la familia de los *modelos logísticos*. Estos últimos son más tratables desde el punto de vista computacional, ya que los normales se definen a través de integrales. Por el contrario, el modelo logístico viene descrito por una función explícita entre los parámetros del ítem y el rasgo latente, y además tiene importantes propiedades estadísticas. Otra ventaja de utilizar la función logística radica en que los parámetros que definen el modelo normal, en el modelo logístico siguen manteniendo su interpretación gráfica.

En la actualidad, los modelos logísticos son los más populares y pueden ser clasificados a su vez en función del número de parámetros que los caracterizan. Como resultado existen modelos logísticos de un parámetro (1PL), conocido también como modelo de Rasch (1960), de dos parámetros (2PL) y de tres parámetros (3PL), estos dos últimos propuestos por Birnbaum (1968). Genéricamente, la CCI logística viene definida por la siguiente función:

$$P(\theta) = P(u|\theta) = c + (1 - c) \frac{1}{1 + e^{-Da(\theta-b)}} \quad (2.11)$$

donde D es un factor de escala que se introduce para que la función se asemeje lo más posible a la función normal, manteniéndose de esta forma sus propiedades. Teóricamente, θ toma valores reales entre $-\infty$ y ∞ , aunque en la práctica se consideran valores entre -4 y 4 . Los tres parámetros restantes son:

- *Factor de discriminación (a_i)*: Es un valor proporcional a la pendiente de la curva. Un valor alto indica que la probabilidad de que un individuo con un rasgo latente estimado mayor que la dificultad del ítem acierte es mayor. Cuanto más discriminante es un ítem mejor contribuye a una estimación más precisa del conocimiento del alumno.
- *Dificultad (b_i)*: Corresponde al valor de θ para el cual la probabilidad de responder correctamente al ítem es la misma que de responder de forma incorrecta, descontando el factor de adivinanza. Analíticamente representa el grado en el que la curva

está desplazada a la izquierda (tendencia a la facilidad) o a la derecha (tendencia a la dificultad) (Olea y Ponsoda, 2001) con respecto al eje de abscisas.

- *Factor de adivinanza* (o pseudoazar)(c_i): Es la probabilidad de que un alumno sin conocimiento ninguno responda correctamente a la pregunta. Mediante este parámetro el modelo contempla el caso en el que un examinando acierta, habiendo respondido de forma aleatoria.

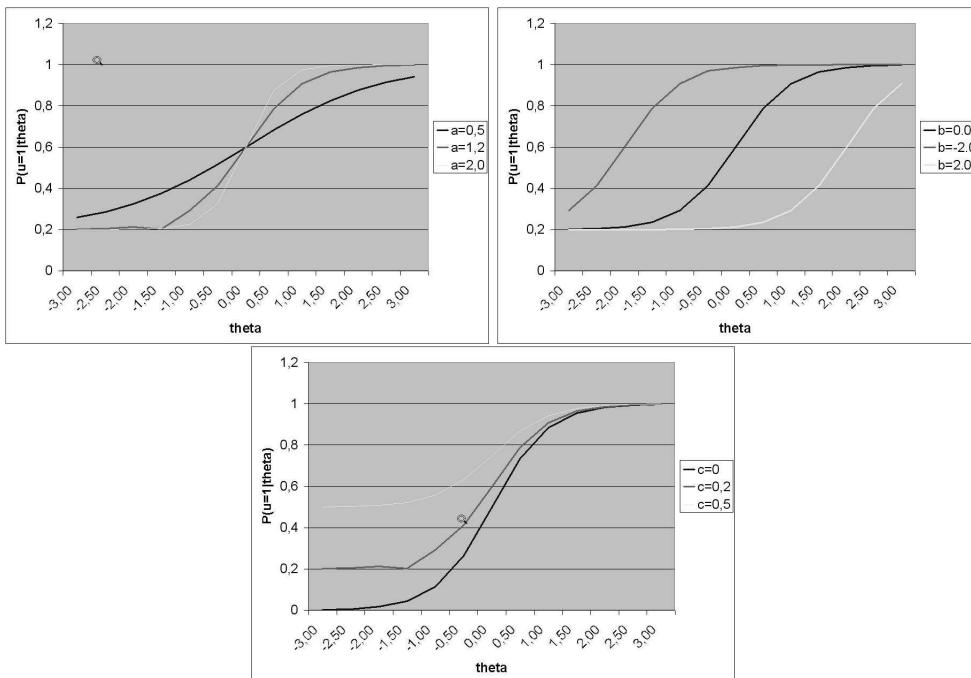


Figura 2.2: Representación gráfica de la función 3PL.

En las gráficas de la figura 2.2 se puede apreciar fácilmente la influencia de los parámetros de la función 3PL en la forma de la propia curva. La primera de ellas representa la variación con respecto al factor de discriminación. Obsérvese que cuanto mayor es su valor, mayor es la probabilidad de que alumnos con niveles de conocimientos altos respondan correctamente al ítem, y a su vez menor es la probabilidad de que examinandos con bajos niveles de conocimiento acierten. La segunda gráfica muestra la variación con respecto a la dificultad. Se puede constatar fácilmente que el cambio en su valor supone un desplazamiento horizontal de la curva. Cuanto mayor es la dificultad más separado está el punto de inflexión de la curva del eje de ordenadas. La tercera gráfica describe cómo afecta la variación en el factor de adivinanza a la forma de la CCI. En este caso, cuanto mayor es su valor más se aleja la curva del eje de abscisas en los niveles de conocimientos más bajos. Esto indica que cuanto mayor sea el valor de la adivinanza, mayor es la probabilidad de que los alumnos de niveles de conocimiento más bajos respondan correctamente. Por último, si en la función 3PL se asume que el factor de adivinanza es igual a cero, la ecuación que se obtiene es la que describe al modelo 2PL. Si además el factor de discriminación es igual a uno, la ecuación resultante describe al modelo 1PL.

Birnbaum fue el primero en reemplazar el modelo normal por uno logístico, motivado por los resultados obtenidos por Haley (1952), según los cuales, para una función de distribución logística $L(x)$ (donde la discriminación y dificultad son iguales a uno, y la adivinanza igual a cero) con factor de escala $L(1,7x)$, y una función de distribución normal $N(x)$ (con media cero y varianza uno), se cumple que:

$$|N(x) - L(1,7x)| < 0,01 \quad \forall x \in (-\infty, \infty) \quad (2.12)$$

Como consecuencia, y para mantener las propiedades que caracterizan a los modelos normales, en la formulación de los modelos logísticos D toma el valor 1,7.

En los modelos anteriores, la CCI se ajusta a una familia de funciones. Es decir, la CCI queda completamente descrita a partir de un conjunto de parámetros. Por este motivo, a éstos se les denominan *modelos paramétricos*. Existen otro tipo de modelos en los que las CCI no pertenecen a ninguna familia, sino que vienen definidas por un conjunto de valores asociados al rango que puede tomar el rasgo latente. Por consiguiente, estos modelos responden únicamente a datos estadísticos obtenidos a partir de la calibración de las curvas con datos reales, y se denominan *modelos no paramétricos* (Sijtsma y Molenaar, 2002). Éstos se estudiarán con más detalle en la sección 2.8.4.

Clasificación basada en el tratamiento de la respuesta

Otro criterio de clasificación de modelos de la TRI se basa en la forma en la que se utiliza la respuesta dada por el alumno. Según este criterio los modelos pueden clasificarse en:

- *Dicotómicos o binarios*: Engloban a la mayoría de los modelos de la TRI, y sólo tienen en cuenta si la respuesta proporcionada por el examinando ha sido correcta o incorrecta. No tienen en cuenta, por tanto, qué respuesta en concreto ha sido seleccionada.
- *Polítómicos*: Tienen en cuenta cuál ha sido la respuesta seleccionada por el alumno, a la hora de llevar a cabo la actualización de su estimación de conocimiento. De esta forma, cada respuesta tiene asociada una curva característica que expresa la probabilidad de que un examinando, con un nivel de conocimiento dado, seleccione esa respuesta. Estos modelos se utilizan principalmente en tests de actitud y de personalidad. Serán tratados con mayor detenimiento en la sección 2.8.2.

Taxonomía de Thissen y Steinberg

Por último, y fuera ya de los criterios definidos en esta sección, se ha querido reseñar una de las taxonomías de clasificación más populares, la propuesta por Thissen y Steinberg (1986) y citada en (Dodd et al., 1995). En esta taxonomía, la categorización se lleva a cabo en función de las restricciones de los parámetros de los modelos. Además, presenta la desventaja de que no se tienen en cuenta los modelos no paramétricos ni los multidimensionales. En cualquier caso, según esta taxonomía, mostrada en la figura 2.3, los modelos pueden ser de cinco tipos, tal y como se enumera a continuación:

- *Modelos binarios*: (en inglés, *Binary models*). Corresponden a aquellos en los que la respuesta del alumno sólo puede ser clasificada como correcta o incorrecta. Son los

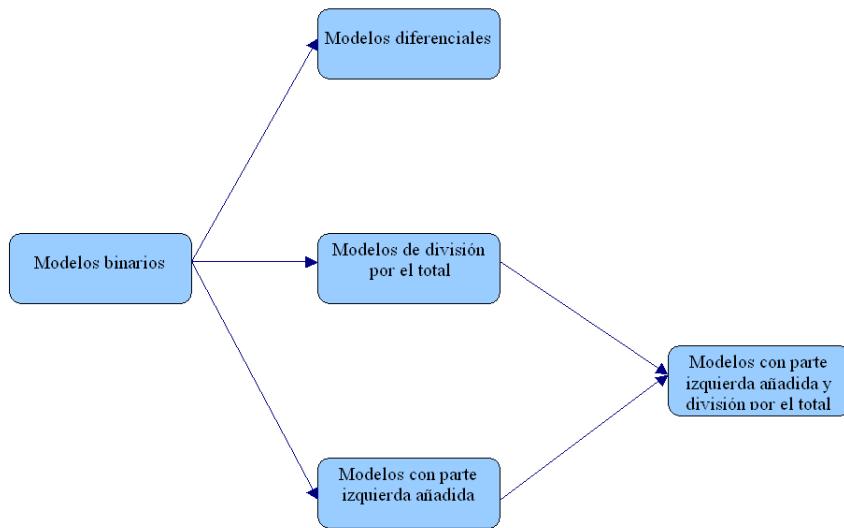


Figura 2.3: Taxonomía de modelos basados en la TRI.

primeros modelos de la TRI, y han servido de base para la construcción de muchos de los posteriores. Son por ejemplo, el modelo de Lord (1980), el de Rasch (1960), y el 2PL de Birnbaum (1968).

- *Modelos diferenciales*: (en inglés, *Difference models*). Dentro de este grupo se encuentran el modelo de escala de clasificación de Muraki (en inglés, *Muraki's rating scale model*) (Muraki, 1990), y el *de respuesta graduada* (en inglés, *graded response model*) (Samejima, 1969, 1997). Éste último utiliza como base los modelos binarios normales y logísticos, y en él se calculan curvas características asociadas a cada respuesta, a partir de diferencias entre funciones logísticas. En la sección 2.8.3 se explicarán con mayor nivel de detalle sus características.
- *Modelos de división por el total*: (en inglés, *Divide-by-total models*). Esta categoría engloba a modelos politómicos en los que las curvas características de cada respuesta tienen la peculiaridad de que se componen de un numerador que caracteriza a cada respuesta, y de un denominador que es igual a la suma de los numeradores de las curvas de todas las posibles respuestas. Dentro de este grupo se encuentran el *modelo de respuesta nominal* (en inglés, *nominal response model*) (Bock, 1972, 1997), el *de crédito parcial generalizado* (en inglés, *generalized partial credit model*) (Muraki, 1992, 1997), el *de crédito parcial* (en inglés, *partial credit model*) (Masters, 1982; Masters y Wright, 1997), y el *de escala de clasificación* (en inglés, *rating scale model*) (Andrich, 1978).
- *Modelos con parte izquierda añadida*: (en inglés, *Left-side added models*). Estos modelos tienen en cuenta la posibilidad de que un examinando, sin conocimiento ninguno de la materia evaluada, responda al ítem seleccionando una respuesta al azar. Son por tanto, modelos en los que se introduce el denominado factor de adivinanza. El más característico de esta categoría es el 3PL de Birnbaum (1968).

- *Modelos con parte izquierda añadida y división por el total:* (en inglés, *Left-side added divide-by-total models*). Combinan las características de las dos categorías anteriores. Dentro de este nuevo grupo se encuentra el modelo de opción múltiple de Thissen y Steinberg (1984).

2.8.2. Modelos politómicos

En los modelos dicotómicos los ítems se evalúan únicamente en dos categorías: correctos o incorrectos. Por el contrario, los politómicos permiten clasificar la respuesta del alumno en diversas categorías. De esta forma, cada respuesta es tratada de forma independiente. Según Embretson y Reise (2000), diversos investigadores han puesto de manifiesto que los modelos de la TRI politómicos son más informativos que los dicotómicos. Como consecuencia, se pueden llegar a obtener estimaciones más precisas con un número más reducido de ítems (Hontangas et al., 2000).

Hasta la década de los 70, los modelos de la TRI utilizados eran dicotómicos (van der Linden y Hambleton, 1997). Cuando los datos obtenidos eran politómicos (por ejemplo, la puntuación obtenida en un ítem de redacción), éstos eran dicotomizados (Drasgow, 1995). Los primeros modelos politómicos datan de finales de la década de los 60, y aunque durante las décadas de los 70 y los 80 se desarrollaron diversas propuestas en esta línea, no fue hasta la aparición de la herramienta de calibración MULTILOG (Thissen, 1988) y de los PC, cuando éstos fueron realmente abordables. El objetivo con el que fueron inicialmente concebidos era modelar ítems de opción múltiple, de forma que se constatara la influencia en la respuesta proporcionada por el examinando, más que en si ésta ha sido correcta o incorrecta. A pesar de su objetivo inicial, en la actualidad estos modelos se aplican principalmente a ítems en los que las respuestas están ordenadas en categorías.

En los tests de aptitud y en los de personalidad usualmente se incluyen ítems con respuestas ordenadas en categorías denominadas escalas de tipo Likert (o escalas sumativas). Es decir, escalas (en general entre 1 y 5, o entre 1 y 7) en las que el alumno muestra el grado en que está de acuerdo con una determinada sentencia. La Dra. Samejima (1969) introdujo el *modelo de respuesta graduada* (en inglés, *graded response model*). Esta propuesta y sus posteriores variaciones representan el primer modelo politómico desarrollado, y además en el que se han inspirado otros posteriores.

Como se verá a continuación, la información sobre la respuesta seleccionada (en vez de considerar si es correcta o incorrecta) puede llegar a ser bastante relevante para evaluar a un individuo de forma eficiente. Para ilustrar su importancia, se utilizará un ejemplo similar al citado en (Thissen y Steinberg, 1997).

Supongamos el ítem de la figura 2.4. Se trata de una pregunta similar a otra utilizada en un test de evaluación administrado a alumnos de tercer y cuarto grado en Carolina del Norte (EE.UU.). Una vez suministrado el test, se comprobó que los alumnos con mayores niveles de conocimiento respondían correctamente a este ítem seleccionando la opción B. La mayoría de los examinandos con niveles de conocimiento bajos elegían la alternativa A o la D. Asimismo, los resultados mostraron que un amplio número de sujetos con un nivel de conocimiento relativamente alto marcaban la opción C. Realmente aunque la opción C es incorrecta, su grado de incorrección es menor que el de las opciones A y D. Este aspecto no se tiene en cuenta en los modelos dicotómicos, los cuales consideran igualmente erróneas las respuestas A, C y D.

Además, los modelos politómicos permiten dar lo que en inglés se denomina *crédito parcial* (*partial credit*) a la respuesta que un alumno da a un ítem. En dominios como la

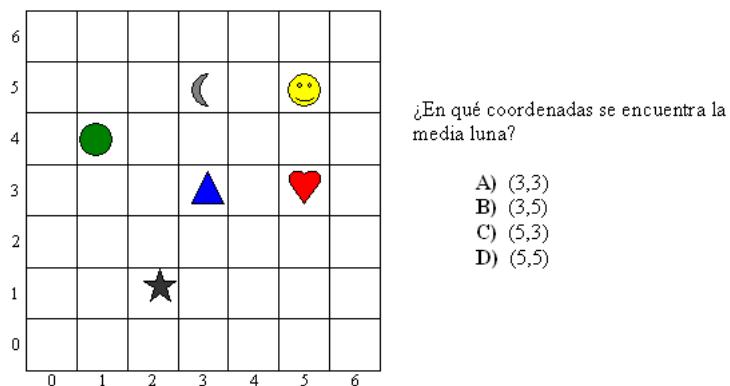


Figura 2.4: Ítem de opción múltiple

Física, la Química o las Matemáticas puede ser adecuado el uso de ejercicios complejos en los que el examinando debe resolver un conjunto de pasos para obtener una respuesta totalmente correcta (ítems por partes). A través de los modelos politómicos, es posible dar una calificación parcialmente correcta a aquellos estudiantes que hayan desarrollado correctamente algunos pasos.

Según la taxonomía de Thissen y Steinberg (1986) vista anteriormente, los modelos politómicos pueden clasificarse en tres tipos: *modelos de diferencia, con división por el total, y con división por el total y parte izquierda añadida*. En el ámbito de los modelos politómicos, las dos primeras categorías hacen referencia a modelos con respuestas categorizadas, es decir, ordinales. Los modelos de diferencias se caracterizan porque, para obtener la curva característica de cada categoría o respuesta, se lleva a cabo una resta con respecto a la categoría anterior o posterior.

Otra posible clasificación de los modelos politómicos atiende a si las respuestas están ordenadas, o si por el contrario el orden de éstas no es relevante. Atendiendo a este criterio, pueden clasificarse en dos tipos (Mellenbergh, 1995): *ordinales o nominales*. En los primeros, las respuestas están ordenadas entre sí. Son la gran mayoría de los modelos politómicos, y de hecho los más populares. En los modelos de la segunda familia, las respuestas no presentan ningún ordenación predeterminada. En este caso cada respuesta es independiente. El objetivo es extraer de los ítems más información sobre el nivel de conocimiento del examinando que cuando se aplican modelos dicotómicos (van der Linden y Hambleton, 1997). Esta familia de modelos están más estrechamente relacionados con el modelo presentado en este trabajo, por este motivo, en la siguiente sección se estudiarán algunas de las propuestas más relevantes en este ámbito.

En general, en los modelos de respuesta politómicos, además de la CCI, se define una curva característica por cada respuesta posible al ítem. Estas curvas son denominadas en la literatura *líneas de traza* (en inglés, *trace lines*), funciones características de operación (en inglés, *operating characteristics functions*) (Dodd et al., 1995). A lo largo de este trabajo, esas curvas serán denominadas *Curvas Características de Respuesta* (CCR). Éstas, análogamente a las CCI, sirven para caracterizar cada respuesta del ítem, y al igual que las CCI deben ser estimadas. Pueden describirse como la porción de una población de alumnos con un determinado nivel de conocimiento que, cuando se les administra el ítem, seleccionan esa respuesta.

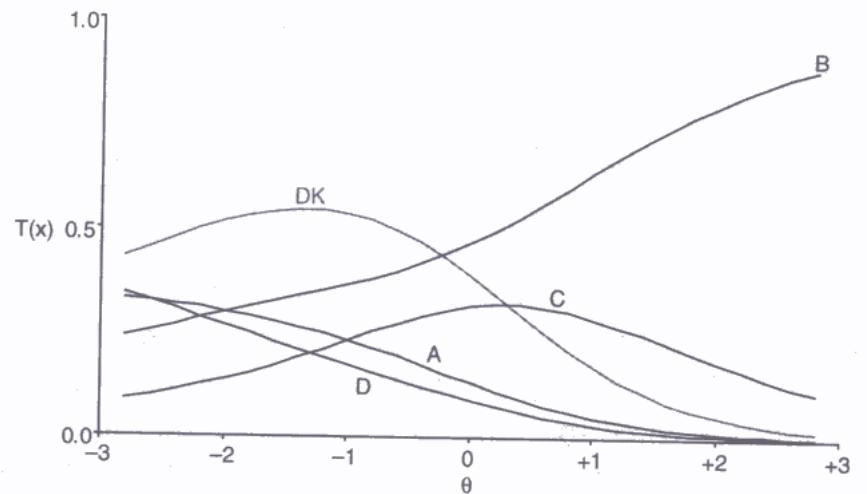


Figura 2.5: Representación de las CCR de un ítem. Extraído de (Thissen y Steinberg, 1997)

La figura 2.5 representa las CCR calibradas del ítem de opción múltiple de la figura 2.4, según el modelo de respuesta para ítems de opción múltiple (Thissen y Steinberg, 1997). La respuesta correcta corresponde a la respuesta *B*. Esto es fácilmente apreciable porque la curva característica es monótona creciente, es decir, cuanto mayor es el nivel de conocimiento del alumno θ mayor es la probabilidad de que seleccione ese respuesta. El resto de opciones se suelen denominar *distractores*. Las curvas de las opciones *A* y *D* son claramente erróneas. Éstas son monótonas decrecientes, es decir, cuanto mayor es el nivel de conocimiento del alumno θ menor es la probabilidad de que el alumno seleccione ese respuesta. La respuesta *C* correspondería a una opción que no siendo correcta es bastante similar a ella. Este tipo de respuestas permiten discriminar a los estudiantes con altos niveles de conocimiento con respecto a los de niveles medios. La CCR de este tipo de respuestas es no monótona, y se puede apreciar que aquellos individuos con niveles de conocimiento intermedios tienen mayor probabilidad de seleccionarla. Por último, la curva *DK* modela la situación de aquellos examinandos que han dejado la respuesta en blanco. Como se verá en la siguiente sección, algunos modelos politómicos tienen en cuenta esta categoría, que se denomina comúnmente *respuesta latente* (en inglés, *latent response*) o *respuesta desconocida* (en inglés, *don't know*), como una alternativa más.

Como se ha mencionado anteriormente, la ventaja principal del uso de modelos politómicos es que éstos son más informativos. Por este motivo, para poder suministrar un TAI con un modelo de respuesta de este tipo, el tamaño exigido a un banco de ítems se reduce bastante. Dodd et al. (1995) realizaron un estudio de diversos modelos politómicos aplicados a TAI, del que se extrajeron diversas características que tienen en común este tipos de TAI. La más destacada es acerca del tamaño del banco de ítems. Los TAI con modelos dicotómicos requieren un banco de ítems de al menos 100 ítems, o de entre 500 y 1000 para tests adaptativos de longitud fija, contenido balanceado y en los que la seguridad de los ítems es un aspecto importante. Por el contrario, este trabajo mostró que el tamaño del banco para modelos politómicos es menor.

La principal desventaja de los modelos politómicos radica en la dificultad para estimar los parámetros de las CCR de forma correcta, ya que es necesario disponer de grandes

muestras para la calibración. Si ya de por sí es costoso estimar las CCI para los modelos dicotómicos, el problema se agrava en los politómicos ya que el número de curvas a estimar se multiplica por el número de respuestas de cada ítem. Estudios de simulación realizados han determinado que para poder estimar correctamente los parámetros de una CCR son necesarios de uno a cinco sujetos por parámetro (Hontangas et al., 2000). Éste es quizás el principal motivo por el que este tipo de modelos, en la práctica, no suelen ser utilizados en TAI.

2.8.3. Los modelos de respuesta politómicos basados en la TRI

En esta sección se recogen las características principales de algunos de los modelos politómicos más populares.

Modelo de respuesta graduada

Este modelo (Samejima, 1969, 1997) se utiliza en tests evaluados en categorías ordenadas del tipo Likert. Es una generalización del modelo 2PL, en la que se permite que los ítems tengan un número diferente de categorías de respuesta. Se caracteriza porque la probabilidad condicional de que un alumno responda una determinada categoría se calcula en dos pasos. Supóngase un ítem i con $m_i + 1$ posibles respuestas. Asociadas a éstas se consideran m_i umbrales entre cada dos consecutivas. En el primer paso, se calculan m_i curvas, una por cada umbral, de la siguiente forma:

Pregunta: ¿Me encanta asistir a fiestas muy concurrencias?



Figura 2.6: Ítem politómico con categorías ordenadas (adaptado de (Embretson y Reise, 2000)).

$$P_{ij}^*(\theta) = \frac{e^{\alpha_i(\theta - \beta_{ij})}}{1 + e^{\alpha_i(\theta - \beta_{ij})}} \quad (2.13)$$

donde $j = 1, \dots, m_i$. Estas curvas son denominadas *Curvas Características de Funcionamiento*(CCF) (en inglés, *Operating Characteristic Curves*), y representan la probabilidad de que la respuesta elegida por un alumno sea mayor o igual que la categoría correspondiente. Los parámetros β_{ij} tienen la siguiente interpretación: indican el rasgo latente necesario para responder por encima del umbral j con una probabilidad de 0,5. Es decir, el ítem es tratado como m_i ítems dicotómicos (la respuesta 0 frente a la 1,2,...; la 0,1 frente a la 2,3,..., etc.). De esta forma, por cada uno de ellos se estiman todas sus CCF, con la restricción de que todas ellas deben tener la misma pendiente α_i . Una vez que las CCF de los umbrales

han sido computadas, en el segundo paso, se calculan las curvas de las respuestas de la siguiente forma:

$$P_{ij}(\theta) = P_{ij}^*(\theta) - P_{i(j+1)}^*(\theta) \quad (2.14)$$

donde la probabilidad de seleccionar una respuesta mayor o igual que la menor de las respuestas es $P_{i0}^*(\theta) = 1$, y la probabilidad de seleccionar una mayor que la menor de las categorías es $P_{im_i}^*(\theta) = 0$. Estas últimas reciben el nombre de *Curvas de Respuesta Categóricas* (CRC) (en inglés, *Category Response Curve*).

En este modelo, los parámetros de las curvas determinan la forma y la localización de las CRC y las CCF. Cuanto mayor es el valor de la pendiente α_i , más pronunciada es la CCF, y más estrecha y apuntada es la CRC, indicando que es muy fácil distinguir entre las distintas categorías de respuesta. Los parámetros β_{ij} de cada umbral de categorías determinan la localización de las CCF, así como el mayor valor de la CRC.

Modelo de respuesta nominal

Propuesto por Bock (1972, 1997), permite modelar ítems cuyas respuestas no están necesariamente ordenadas (Embretson y Reise, 2000). Todos los modelos politómicos de la familia de los con división por el total pueden considerarse casos especiales de éste (Thissen y Steinberg, 1986). Su objetivo inicial era poder representar las respuesta distractoras mediante curvas características.

En esta propuesta la probabilidad de que un examinando responda seleccionado la respuesta o categoría x , siendo $x = 0...m_i$, puede expresarse de la siguiente forma:

$$P_{ix}(\theta) = \frac{e^{\alpha_{ix}\theta + c_{ix}}}{\sum_{x=0}^m e^{\alpha_{ix}\theta + c_{ix}}} \quad (2.15)$$

Donde se imponen la restricción: $\sum \alpha_{ix} = \sum c_{ix} = 0$, o para el caso de categorías ordenadas $\alpha_{i1} = c_{i1} = 0$. Los parámetros α_{ix} y c_{ix} tienen que ser estimados para cada una de las $m_i + 1$ categorías de cada ítem i . α_{ix} representa la discriminación de la CCR de la respuesta x , y c_{ix} es lo que se denomina *parámetro de intercepción* (en inglés, *intercept parameter*), cuyo significado parece no estar muy claro.

Modelo de crédito parcial generalizado

Esta propuesta (Muraki, 1992), denominada inicialmente *modelo de crédito parcial* (Masters, 1982), fue concebida para tratar ítems para cuya resolución es necesario realizar un conjunto de pasos (ítems por partes), como por ejemplo, los aplicados en la resolución de problemas matemáticos. La evaluación de éstos debe realizarse en función del número de pasos llevados a cabo correctamente. De esta forma, aunque no se haya completado el ítem adecuadamente, debe concederse al examinando lo que se denomina "crédito parcial", que tenga en consideración los pasos realizados de forma correcta.

Se trata de un modelo de división por el total, cuya primera versión se basaba en el dicotómico 1PL. Sin embargo, su versión generalizada incorpora un factor de discriminación a la formulación original, con lo que se puede considerar como una extensión del modelo dicotómico 2PL. Dado un ítem i con $K_i + 1$ categorías de respuestas, $x = 0, ..., m_i$, la curva característica de la respuesta viene definida de la siguiente forma:

$$P_{ix}(\theta) = \frac{e^{\sum_{j=0}^x \alpha_i(\theta - \delta_{ij})}}{\sum_{r=0}^{m_i} [e^{\sum_{j=0}^r \alpha_i(\theta - \delta_{ij})}]} \quad (2.16)$$

donde $\sum_{j=0}^0 \alpha_i(\theta - \delta_{ij}) = 0$. El parámetro δ_{ij} se denomina *dificultad del paso* asociado a la categoría j . Cuanto mayor sea su valor, más difícil es ese paso con respecto a los otros. Analíticamente, este parámetro corresponde al rasgo latente en el que se encuentra la intersección entre dos categorías consecutivas. Por este motivo, δ_{ij} se denomina también *intersección de categorías*. La pendiente α_i no tiene la misma interpretación que en el modelo 2PL. En esta ocasión indica el grado con el cual las respuestas varían entre ítems con respecto a las variaciones de θ .

Modelo de respuesta para ítems de opción múltiple de Thissen y Steinberg

Esta propuesta realizada por Thissen y Steinberg (1997) es una combinación de las ideas aportadas por los dos modelos anteriores. En ella, las curvas de las respuestas vienen caracterizadas de la siguiente forma:

$$P_{ix} = \frac{e^{a_{ix}\theta + c_{ix}} + d_{ix}e^{a_{i0}\theta + c_0}}{\sum_{k=0}^m e^{a_{ik}\theta + c_k}} \quad (2.17)$$

Como se puede apreciar, la ecuación 2.17 es idéntica a la propuesta por Bock en la ecuación 2.15, pero añadiendo un segundo sumando en el numerador. Esto es debido a que, en este caso, se tiene en cuenta la respuesta o categoría latente (respuesta en blanco), cuya curva de respuesta viene descrita por la siguiente ecuación:

$$P_{i0} = \frac{e^{a_{i0}\theta + c_0}}{\sum_{k=0}^m e^{a_{ik}\theta + c_k}} \quad (2.18)$$

A la vista de las ecuaciones 2.17 y 2.18, se puede observar que este modelo, en cada categoría, añade a la probabilidad de que el alumno responda seleccionando la respuesta correspondiente, un cierto porcentaje d_{ix} de la curva de la respuesta en blanco.

El parámetro a_{ik} representa el poder discriminativo de la respuesta. Si su valor es grande, la curva será monótona creciente. Si por el contrario es pequeño, será monótona decreciente. Para valores intermedios, la curva no es monótona. Para otra parte, el parámetro c_{ik} indica la frecuencia relativa de selección de cada respuesta. Para respuestas cuyo parámetro a_{ik} es similar, la de mayor valor de c_{ik} será la más seleccionada. Al igual que en el modelo de Bock, se impone sobre los parámetros de las curvas la restricción $\sum \alpha_{ix} = \sum c_{ix} = 0$. Por último, d_{ix} es la proporción de individuos que no sabiendo la respuesta, seleccionaron la opción x . Como se trata de un porcentaje, debe satisfacerse la siguiente condición: $\sum_{k=1}^{m_i} d_{ik} = 1$.

Modelos basados en el proceso de respuesta de Nedelsky

Existen diversos modelos politómicos que se basan en el denominado proceso de respuesta de Nedelsky (1954), por el cual, el alumno responde a un ítem siguiendo un procedimiento que consta de dos fases: En la primera, descarta aquellas respuestas que considera incorrectas. Posteriormente, selecciona aleatoriamente entre una de las respuestas que quedan.

A continuación se describirán brevemente algunos de los modelos politómicos basados en esta teoría.

Modelo *Full Nedelsky*

Este modelo (Verstralen, 1997a) es similar a otro denominado *Chain Nedelsky* (Verstralen, 1997b) en el que se hacían ciertas suposiciones sobre el procedimiento empleado por el alumno para seleccionar la respuesta correcta a un ítem. Según ambas propuestas, el examinando ordena las opciones según lo fácil que para él es indicar que son erróneas. Así, por ejemplo, para un ítem de cuatro opciones de respuesta, el patrón 0123 indicaría que la opción 0 es la que se ve más claramente que es errónea, y la opción 3 sería la identificada como correcta. Se asume además, que entre el subconjunto de opciones preseleccionadas por los alumnos, siempre está la que es verdaderamente correcta.

Genéricamente, sea x_i la respuesta a un ítem de opción múltiple i con $J_i + 1$ opciones, $x_i \in 0, \dots, J_i$, donde la respuesta correcta es la opción 0, la probabilidad de que $x_i = j$ se puede obtener en dos pasos. En el primero de ellos se calcula:

$$z_{ij} = z_{ij}(\theta) = e^{-a_j\theta + \zeta_{ij}} \quad (2.19)$$

donde $j = 1, \dots, J_i$, y a_i representa el factor de discriminación (positivo), y ζ_{ij} el parámetro de localización. La probabilidad de que una opción incorrecta forme parte del subconjunto de alternativas (α), de entre las cuales un alumno con nivel de conocimiento θ hace su elección, es la siguiente:

$$p_{ij} = p_{ij}(\theta) = \frac{z_{ij}(\theta)}{1 + z_{ij}(\theta)} \quad (2.20)$$

$p_{i0} = 1$ es la probabilidad de que la opción correcta pertenezca al subconjunto de alternativas. Si se asume independencia condicional de éstas dado θ , y se considera que $\alpha^j = 1$ representa que la opción j pertenece a α , la probabilidad de la ocurrencia del subconjunto α es la siguiente:

$$p_{i\alpha} = p_{i\alpha}(\theta) = \prod_{j=1}^{J_i} p_{ij}^{\alpha^j} (1 - p_{ij})^{1 - \alpha^j} \quad (2.21)$$

La probabilidad condicional de responder al ítem i eligiendo la opción j , dado el subconjunto α , viene dada por:

$$P_{i\alpha j} = P_{i\alpha}(j) = P_i(j|\alpha) = \frac{\alpha^j}{|\alpha|} \quad (2.22)$$

siendo $|\alpha|$ el número de opciones del subconjunto α . De esta forma, la probabilidad de que el examinando seleccione la opción j del ítem i queda de la siguiente forma:

$$f_{ij} = f_{ij}(\theta) = f_i(j|\theta) = \sum_{\alpha} P_{i\alpha j} p_{\alpha} \quad (2.23)$$

donde α , en el sumatorio, representa un posible subconjunto de opciones de respuesta erróneas y la correcta 0.

El factor de discriminación α_i podría, o bien ser estimado como en el modelo de crédito parcial generalizado, o bien ser un valor constante conocido. Asimismo, es posible considerar un factor de discriminación diferente α_{ij} por cada opción de respuesta. Sin embargo, esta aproximación dificulta en gran medida la calibración de las curvas de respuesta, requiriendo un número considerable de sesiones de tests.

Modelo de Revuelta para ítems de opción múltiple

En este modelo (Revuelta, 2000), cada ítem i tiene asociado un vector de parámetros ϵ_i , que contiene dos de ellos, α_i y β_i , por cada alternativa incorrecta del ítem. Supongamos que i posee K posibles respuestas. Este ítem podría dividirse en $K - 1$ subítems. Cada uno de ellos contendría el enunciado del ítem original, y sólo dos respuesta: la correcta y una de las incorrectas. Sean $P_{ib}(Y = r|\theta, \epsilon_i)$ y $P_{ib}(Y = o|\theta, \epsilon_i)$ las probabilidades de seleccionar la respuesta correcta (o) y de seleccionar la respuesta incorrecta (r) (en cada subítem) respectivamente. En esta situación, es posible definir el ratio $P^*(i, r, \theta, \epsilon_i)$ de esas dos probabilidades, de la siguiente forma:

$$P^*(i, r, \theta, \epsilon_i) = \frac{P_{ib}(Y = r|\theta, \epsilon_i)}{P_{ib}(Y = o|\theta, \epsilon_i)} \quad (2.24)$$

Este ratio es monótono decreciente en θ . Para definirlo se ha utilizado la función logística 2PL, donde β_r representa la dificultad. Cuanto mayor es su valor, es necesario tener un nivel de conocimiento mayor para poder descartar esa respuesta incorrecta. Igualmente, el parámetro α_r , la discriminación de la curva, en este caso representa el gradiente de la función ratio con respecto a θ .

Para obtener las probabilidades de cada respuesta se asume la regla de elección de Luce (1959), según la cual:

$$\frac{P_i(Y = r|\theta, \epsilon_i)}{P_i(Y = o|\theta, \epsilon_i)} = \frac{P_{ib}(Y = r|\theta, \epsilon_i)}{P_{ib}(Y = o|\theta, \epsilon_i)} \quad (2.25)$$

Asumiendo la igualdad anterior para cada una de las respuestas incorrectas y la restricción $\sum_{a=1}^A P_i(Y = a|\theta, \epsilon_i) = 1$, se obtiene un sistema de ecuaciones. Resolviéndolo se infieren las probabilidades de cada respuesta, que se pueden expresar genéricamente de la siguiente forma:

$$P_i(Y = r|\theta, \epsilon_i) = \frac{P^*(i, r, \theta, \epsilon_i)}{\sum_{a=1}^K P^*(i, a, \theta, \epsilon_i)} \quad (2.26)$$

2.8.4. Modelos no paramétricos

Aunque los modelos no paramétricos se llevan utilizando desde hace unos cincuenta años, no fue hasta principios de los ochenta cuando se reavivó su (van der Linden y Hambleton, 1997, part. IV). Los *modelos no paramétricos* (en inglés, *non-parametric models*) son una familia de modelos basados en la TRI, en los que se delimita cuál es el conjunto mínimo de suposiciones que deben cumplirse para obtener medidas válidas sobre personas e ítems (Sijtsma y Molenaar, 2002). Se utilizan sobre todo en tests de personalidad y de aptitud.

La aplicación de modelos paramétricos implica asumir un conjunto de supuestos que no siempre son correctos, a la vista de los datos reales utilizados para llevar a cabo la calibración de los ítems. Existen determinadas situaciones en las que las funciones que describen las curvas características, no son adecuadas para describir el comportamiento real de un alumno cuando responde a un determinado ítem. Este problema se tratará con más detalle en el capítulo 4. Por otra parte, según Mokken (1997) hay situaciones en las que la información de la que se dispone a priori para llevar a cabo la calibración de los ítems, o para verificar que

sus parámetros son correctos, es insuficiente. En todos estos casos, no es adecuado utilizar modelos paramétricos, y la solución es utilizar los no paramétricos.

En cuanto a las características que deben poseer los modelos no paramétricos basados en la TRI, a las ya comentadas (invarianza, independencia local y unidimensionalidad) con anterioridad, se suman las siguientes, que hacen referencia a características que deben tener las CCI:

- *Monotonidad de las CCI*, es decir que las curvas características sean monótonas crecientes en θ .
- *No intersección de las CCI*, esto es, que no exista intersección alguna a lo largo de θ , entre las distintas CCI del banco de ítems. Esto a su vez implica que las CCI podrán ordenarse y numerarse, tal y como se muestra a continuación:

$$P_1(u|\theta) \leq P_2(u|\theta) \leq \dots \leq P_n(u|\theta), \quad \forall \theta \quad (2.27)$$

En esta línea, las dos propuestas dicotómicas más importantes son (Sijtsma y Molenaar, 2002):

- a) *Modelo de Homogeneidad Monótona* (en inglés, *Monotone Homogeneity Model*): Se utiliza para ordenar a individuos según un cierto criterio medido a través de un test. Este modelo no cumple el requisito de no intersección de las CCI. Se basa en el supuesto de que la ordenación que se obtendrá (a través de su aplicación) va a coincidir con la obtenida considerando la puntuación verdadera de los individuos.
- b) El *Modelo de Monotonidad Doble* (en inglés, *Double Monotonicity Model*): Es un caso especial del anterior, en el que sí se asume el supuesto de no intersección de las CCI.

Al igual que sucede con los paramétricos, también existen aproximaciones polítómicas, tales como las extensiones de los dos modelos anteriores también descritas en (Sijtsma y Molenaar, 2002). En general, la gran mayoría de los modelos no paramétricos polítómicos se aplican sobre ítems con respuestas ordenadas en categorías, orientados por tanto, a tests de personalidad y de aptitud. Entre los modelos no paramétricos polítómicos nominales destaca el propuesto por Abrahamowicz y Ramsay (1992), en el que se da una orientación no paramétrica al modelo para ítems de opción múltiple de Thissen y Steinberg (*op. cit.*).

Para aquéllos que deseen profundizar más en las aproximaciones no paramétricas de la TRI, el artículo de Junker y Sijtsma (2001b) sirve de introducción para un número especial de la revista *Applied Psychological Measurement* dedicado a este tipo de modelos. Asimismo, el libro de Sijtsma y Molenaar (2002) es también un buen manual de iniciación.

Para finalizar esta sección dedicada a la TRI, decir que a pesar de que esta teoría se ha impuesto hoy en día frente a la TCT como fundamento teórico para la evaluación mediante tests, bien es cierto que existen diversos autores como (Nelson, 2003) que consideran que los beneficios en el uso de la TRI frente a su predecesora no son tan desmesurados como cabría esperar. Lawson (1991) compara la TCT con el modelo de Rasch, realizando tres tests con tres muestras de población diferentes. La conclusión que obtuvo es que los resultados

aplicando ambas teorías son bastante similares. Por otro lado, tanto Fan (1998) como Stage (1998) compararon la TCT con diversos modelos de la TRI de forma empírica con grandes grupos de examinandos, obteniendo también resultados muy similares. Por último, Thorndike (1984) resalta que para evaluar a un número elevando de examinandos, los ítems que serán seleccionados mediante procedimientos basados en la TRI no serán muy diferentes de los elegidos mediante criterios basados en la TCT, y los resultados obtenidos en ambos casos tendrán propiedades muy similares.

2.9. Los Tests Adaptativos Informatizados

Durante bastante tiempo los tests que se realizaban para evaluar a un conjunto de alumnos se hacían sobre papel y lápiz. Como se ha puesto de manifiesto en la sección 2.3, el incremento en el uso de la tecnología ha llevado a las instituciones educativas a aplicar estas tecnologías a la evaluación mediante tests, elaborando los denominados TACs. Una versión más sofisticada de los TAC son los *Tests Adaptativos*. La idea de utilizar adaptación en los tests no es nueva. En el test de inteligencia de Binet y Simon (1905), los ítems se clasificaban según la edad mental de los evaluandos. De esta forma, el profesor infería la edad mental de cada examinando a partir de sus respuestas, y asimismo adaptaba los siguientes ítems a las estimaciones que iba realizando (van der Linden y Glas, 2000). La definición del concepto de *Test Adaptativo* más comúnmente usada y más intuitiva es la aportada por Wainer (1990):

La idea fundamental de un test adaptativo es imitar de forma automática el comportamiento de un examinador. Esto es, si un examinador le presenta al alumno un ítem demasiado difícil para él, éste dará una respuesta errónea, y por lo tanto, la siguiente vez, el examinador presentará una pregunta algo más fácil.

Generalmente, un *Test Adaptativo* comenzará presentando una pregunta de nivel medio. Si el examinando la acierta, la siguiente será algo más difícil. Si por el contrario falla, la siguiente será más fácil. Como se puede apreciar, esta idea no es nueva. En los exámenes orales, las preguntas también se ajustan al nivel de conocimiento del alumno (van der Linden y Glas, 2000), con lo que los Tests Adaptativos no son más que un intento de emular un procedimiento de evaluación personalizada.

Un TAI (Wainer, 1990; Olea et al., 1999) es una herramienta de medida administrada a los alumno por medio de un PC, en vez de utilizando el clásico formato de lápiz y papel. En general, en los TAI, los ítems se muestran de uno en uno, y la presentación de cada uno de éstos, así como la decisión de finalizar el test y la evaluación del alumno se llevan a cabo dinámicamente, basándose en las respuestas del examinando. En términos más precisos, un TAI es un algoritmo iterativo que comienza con una estimación inicial del conocimiento del alumno y continúa con los siguientes pasos (el proceso se ilustra en la figura 2.7):

1. Todos los ítems, que no han sido administrados todavía, son analizados para determinar cuál de ellos contribuye en mayor medida a una mejor estimación del conocimiento del alumno.
2. El ítem se muestra al alumno.
3. En función de la respuesta elegida por el examinando, se estima el nuevo nivel de conocimiento de éste.

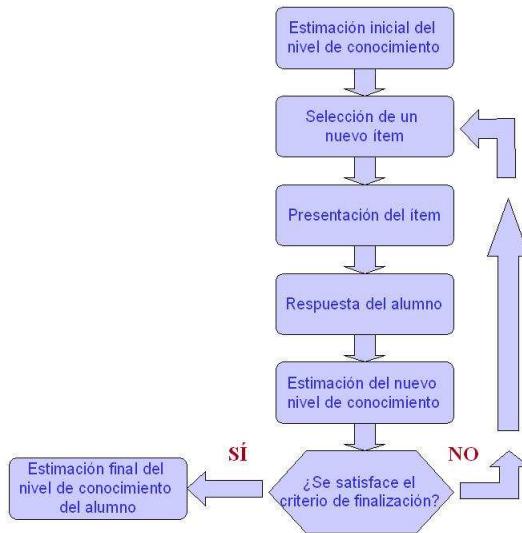


Figura 2.7: Diagrama de flujos de un TAI (adaptado de (Olea y Ponsoda, 2001)).

4. Los pasos del 1 al 3 se repiten hasta que el criterio de finalización del test se satisfaga.

Los criterios para la selección del ítem que debe mostrarse al alumno en cada momento, la decisión de finalizar el test, y la estimación del conocimiento del alumno se basan, en general, en procedimientos bien fundamentados. La selección de ítems y el criterio de parada del test son, por tanto, adaptativos. El número de ítems de un TAI no suele ser fijo, y a cada examinando se le mostrará una secuencia diferente de ítems, e incluso diferentes ítems.

Los elementos principales a la hora de desarrollar un TAI se enumeran a continuación:

- Un *modelo de respuesta* asociado a los ítems: Este modelo describe el comportamiento del alumno en el momento de responder, en función de su nivel de conocimiento estimado. Cuando se desea inferir el nivel de conocimiento, el resultado obtenido debe ser independiente de la herramienta utilizada para llevar a cabo la medición, es decir, ésta debe ser independiente del test utilizado para ello, así como del individuo que lo realiza.
- Un *banco de ítems*: Es uno de los elementos más importantes de un TAI, ya que cuanto mejor sea su calidad, los tests adaptativos serán más precisos. El desarrollo de un buen banco de ítems es la fase más tediosa en la construcción un TAI, ya que éste debe contener un gran número de ítems correctamente calibrados. Una de las restricciones más importantes a tener en cuenta durante el proceso de creación, es que se cumpla el principio de independencia de los ítems. Diversos autores como Flaugher (1990), Barbero (1999) proponen guías de recomendación generales sobre los pasos que deben seguirse en el proceso de desarrollo y construcción de bancos de ítems. En cuanto al número de ítems que debe tener un buen banco, muchas son las recomendaciones, aunque en general, se considera que como mínimo debe estar compuesto de cien ítems.
- El *nivel de conocimiento inicial*: Es muy importante llevar a cabo una buena estimación inicial del nivel de conocimiento del alumno, ya que ésta determinará la duración

final del test. Diferentes criterios pueden utilizarse: la media de los niveles de conocimiento de los alumnos que hayan realizado el test con anterioridad, creación de un perfil y utilizar la media de los alumnos que sean clasificados con ese perfil (Thissen y Mislevy, 1990), etc.

- El *criterio de selección* de ítems: El mecanismo de adaptación propio de los TAI se encarga de seleccionar el siguiente ítem que debe mostrarse al alumno en cada momento. Esta decisión se toma en función del nivel de conocimiento estimado (obtenido a partir de los ítems administrados al alumno con anterioridad). Seleccionar el mejor ítem, es decir, elegir aquel ítem más informativo desde el punto de vista de la estimación, mejora la precisión del test y reduce su número de ítems.
- El *criterio de finalización*: Define la forma en la que se determina la finalización del test. Hay varios modos de hacerlo; el más apropiado, desde el punto de vista de la adaptación, es aquél que finaliza el test cuando la precisión en la estimación del nivel de conocimiento del alumno es mayor que un cierto umbral predefinido. Otros criterios no adaptativos utilizados son: llegar al máximo número de ítems permitidos en un test, haber consumido el tiempo requerido para completar el test, etc.

Además de las ventajas inherentes a la administración por computador, que ya han sido puestas de manifiesto anteriormente en este capítulo, diversos autores (Mislevy y Almond, 1997; van der Linden y Glas, 2000) han indicado las ventajas que aportan los TAI frente a los tests tradicionales realizados con papel y lápiz:

- La principal es que son más eficientes, es decir, el número de ítems requerido para estimar el nivel de conocimiento del alumno es menor, y como consecuencia también el tiempo empleado en su realización.
- Las estimaciones realizadas por un TAI son más precisas y no están sesgadas. Estudios empíricos realizados utilizando alumnos simulados han demostrado que la precisión de la estimación realizada a través de un TAI, es mayor que en un test con el mismo número de ítems, pero en el que éstos son elegidos aleatoriamente. Asimismo estos estudios también concluyeron que, para obtener una estimación del conocimiento del alumno con una precisión determinada, los TAI requieren un número menor de ítems.
- Los TAI se ajustan a las características personales de cada examinando. Esto conlleva también una mejora de la motivación del alumno, ya que las preguntas se adecúan a su nivel de conocimiento. Por este motivo, ni los examinandos se sienten frustrados por la aparición de preguntas difíciles en el test, ni por el contrario se aburren con preguntas excesivamente fáciles.
- Se reduce la ansiedad del alumno durante la realización del test, ya que cada individuo dispone exactamente del tiempo que requiere para su finalización.
- Los tests son más seguros, ya que a cada individuo se le administra un conjunto y un número diferente de ítems.

Todas estas razones hacen que, por regla general, entre un TAI y el mismo test en formato de papel y lápiz, los alumnos prefieran la primera opción (van der Linden y Glas, 2000).

Por el contrario, los TAI presentan algunas desventajas. Una de las más importantes es la seguridad, ya que los examinandos podrían memorizar los ítems de un test y compartir

esta información con sus compañeros. Es por tanto necesario disponer de un gran banco, así como de técnicas para el control de la exposición de los ítems y para la detección de ítems comprometidos.

Otro aspecto importante es que los ítems deben estar bien calibrados. Sin embargo, para llevar a cabo esta tarea, es necesario disponer de un número considerable de sesiones de tests y esto no siempre es posible. Además, sería recomendable que esta calibración se repitiera conforme nuevos estudiantes realicen tests con esos ítems. Por otro lado, algunos alumnos pueden sentirse discriminados frente a otros por el mero hecho de haber respondido a diferentes ítems y haber realizado tests con diferente número de ellos, que el resto de sus compañeros. Otra queja bastante común es la imposibilidad de modificar la respuesta a una pregunta una vez que ha sido respondida (Wainer, 2000).

2.9.1. Aplicación de la TRI a los TAI

El desarrollo de la TRI a mediados de los 50 supuso un espaldarazo definitivo a los TAI, ya que los dotaba de una base teórica con un sólido fundamento. La TRI ha sido aplicada con éxito en los TAI, no sólo para llevar a cabo una estimación cuantitativa del conocimiento del alumno, sino también en la selección de ítems, y en la decisión de finalización de los test. En este dominio, el rasgo latente es el *nivel de conocimiento*. La TRI se utiliza en los tests adaptativos para desempeñar las siguientes funciones (Wainer y Mislevy, 1990):

1. Caracterizar la diferencia entre los ítems de una forma útil, a través de los parámetros que describen sus curvas características. Esto se consigue, cuando están correctamente calibrados.
2. Determinar reglas eficientes para la selección de ítems. Estimando el nivel de conocimiento en cada momento, es posible determinar qué ítem es más informativo, esto es, cuál contribuye mejor a la estimación del conocimiento del alumno.
3. Establecer una escala de medida común, incluso cuando examinandos diferentes hayan realizado tests diferentes en momentos distintos. Si los ítems están bien calibrados, se asegura que la escala es única.

A continuación se procederá a explicar cómo se aplica la TRI a las principales fases en las que se descompone la realización de un TAI y que fueron descritas en la figura 2.7.

Procedimientos de arranque

El primer paso del algoritmo de un TAI es la selección del primer ítem que le va a ser mostrado al examinando. El problema radica en que no se dispone de ninguna información a priori sobre su nivel de conocimiento. Por este motivo, es necesario determinar, de alguna forma, cuál va a ser el nivel de conocimiento del alumno inicialmente. Por regla general, se asume que el valor del que se parte es igual a la media (o la moda) del nivel de la población. Este dato se obtiene a partir de las estimaciones del conocimiento de la muestra de alumnos que han intervenido en la calibración de los ítems del test.

Por otro lado, si se dispusiera de información adicional sobre el examinando, tal como su edad, estudios, etc. podría extraerse de la muestra poblacional inicial, un subconjunto de alumnos con características similares, y calcular la media de conocimiento de ese subconjunto.

Otra alternativa consiste en administrar inicialmente un número reducido de ítems (tres o cuatro) seleccionados de forma aleatoria, y a partir de ese momento, comenzar el proceso adaptativo.

En cualquier caso, el problema de inferir el nivel de conocimiento inicial del alumno en el test, es más importante en tests con un número reducido de ítems, por ejemplo, diez. En tests de un tamaño mayor (más de 20 o 30 ítems) el algoritmo es capaz de recuperarse de una mala inferencia inicial, y hacer una estimación final precisa del conocimiento del alumno (van der Linden y Pashley, 2001).

Estimación del nivel de conocimiento

Una vez que los parámetros de las CCI están correctamente calibrados, el valor de θ puede ser estimado utilizando la respuesta dada por el alumno en cada ítem. Existen varios métodos para obtener este valor. Todos ellos se basan en calcular una curva de distribución que representa la estimación inferida utilizando la TRI. La forma en la que la curva se estima, así como la técnica utilizada para inferir el nivel de conocimiento final (a partir de la curva de distribución), determinan la diferencia entre los métodos.

Uno de los más utilizados es el *método de estimación de máxima verosimilitud* (Lord, 1980), en el que el nivel de conocimiento de un alumno viene descrito por el valor máximo de la función de verosimilitud $L(u|\theta)$, calculada de la siguiente forma:

$$P(\theta|u) = L(u|\theta) = \prod_{i=1}^n P_i(u_i = 1|\theta)^{u_i} (1 - P_i(u_i = 1|\theta))^{1-u_i} \quad (2.28)$$

donde $u = u_1, u_2, \dots, u_n$ es el patrón de respuesta de un alumno a los n ítems del test, y $u_i = 1$, indica que el individuo ha respondido correctamente al ítem administración en la posición i -ésima. Para obtener el máximo de esta función, basta con calcular el valor de θ que hace cero la primera derivada. Para ello, es necesario aplicar algún procedimiento de aproximación. Uno de los más utilizado es el método iterativo de Newton-Raphson (Santisteban y Alvarado, 2001).

La principal limitación de este método está en que es incapaz de proporcionar una estimación del conocimiento del alumno bajo patrones de respuesta constantes, esto es, un patrón de respuestas con ninguna pregunta respondida de forma correcta, o todas respondidas correctamente.

En el *método de estimación bayesiano* de Owen (1969, 1975) se solucionó el problema del procedimiento anterior. En este método, se calcula, aplicando el teorema de Bayes, la distribución de probabilidad a posteriori del conocimiento del alumno ($P(\theta|u)$). Es decir:

$$P(\theta|u) = \frac{L(u|\theta)P(\theta)}{L(u)} \quad (2.29)$$

donde $P(\theta)$ representa la distribución de probabilidades del conocimiento del alumno a priori, que según Baker (2001) es normal. A su vez, $L(u)$ es la verosimilitud del patrón de respuesta. Como se puede apreciar, $L(u)$ no depende de θ , siendo por tanto, constante. Como consecuencia, el proceso de estimación puede ser aproximado al siguiente producto:

$$P(\theta|u) \propto L(u|\theta)P(\theta) \quad (2.30)$$

Cuando $P(\theta)$ es una distribución uniforme, el método bayesiano es equivalente al de máxima verosimilitud.

La aplicación del método bayesiano proporciona como resultado una función de distribución del conocimiento del alumno. Existen dos modelos bayesianos en función de cómo se infiera el nivel de conocimiento del alumno a partir de esa distribución. Cuando se utiliza la moda de la distribución a posteriori como nivel de conocimiento estimado, el modelo se denomina *Estimación bayesiana del Máximo a Posteriori* (MAP). Si el valor que se toma es la media o esperanza matemática, se denomina entonces *Estimación bayesiana de la Esperanza a Posteriori* (EAP) .

El método bayesiano también tiene ciertas limitaciones (Olea y Ponsoda, 2001). La estimación del rasgo latente, aunque idealmente debiera ser independiente, aplicando este criterio de estimación es dependiente de la varianza y la media de la distribución a priori de θ . Dos alumnos con el mismo patrón de respuesta, pero con dos distribuciones de conocimiento a priori distintas pueden llegar a tener calificaciones diferentes. Por otro lado, el error cometido en las estimaciones es inversamente proporcional a la longitud del test. Por último, si se establece un tiempo límite para realizar el test, la existencia de ítems sin responder va a afectar de forma negativa a los alumnos con mayor nivel de conocimiento y positivamente a los de menor nivel. Esto es consecuencia de la tendencia que tiene este criterio a sesgar hacia la media de la distribución a priori.

Mecanismos de selección de ítems

La TRI puede utilizarse también para seleccionar de forma dinámica el siguiente ítem que debe mostrarse al alumno, en función de su nivel de conocimiento actual. Como se mencionó anteriormente, la idea de adaptar la selección de ítems no es original de la TRI. Entre las décadas de los 50 y 70 se utilizaron los llamados *tests de ramificación* (en inglés, *branching tests*) (Hulin et al., 1983; Thissen y Mislevy, 1990). Entre éstos están los denominados *tests en dos estados*, en los que en la mayor parte del test no se realizaban una selección de los ítems adaptada a cada alumno.

Existen otros mecanismos de selección en base a las respuestas previamente registradas. Los *tests de ramificación fija, con estructura en árbol o piramidales*. En todos ellos, los ítems se colocan en un árbol ramificado, y en función de la respuesta del alumno, el ítem elegido es aquél situado en una determinada rama. El procedimiento es el siguiente: si se responde correctamente al ítem de un cierto nodo, el siguiente ítem será el de dificultad mayor, de entre los ítems situados en sus nodos hijos. En caso de que responda de forma errónea, se selecciona un ítem hijo con una dificultad menor. Estos métodos encontraban su justificación en la velocidad de procesamiento de los computadores de la época, en los que una elección puramente adaptativa era impensable.

A partir de los años 70, y como consecuencia del incremento en la capacidad computacional de los equipos informáticos, comenzaron a utilizarse métodos de selección puramente adaptativos. En la actualidad (van der Linden y Hambleton, 1997), los procedimientos adaptativos de elección más populares son:

- *Método de la Máxima Información* (Weiss, 1982). Consiste en seleccionar aquel ítem que maximiza la información en la distribución provisional del conocimiento del alumno hasta el momento. Sea θ_i el nivel de conocimiento actualmente estimado del alumno i , la *función de información* se calcula de la siguiente forma:

$$I_j(\theta_i) = \frac{(P'_j(\theta_i))^2}{P_j(\theta_i)(1 - P_j(\theta_i))} \quad (2.31)$$

siendo $P_j(\theta_i)$ el valor de la CCI para el nivel del alumno, y $P'_j(\theta_i)$ la derivada de la CCI en ese mismo punto. En resumen, el procedimiento seguido es el siguiente: Se calcula el valor de la función de información para todo ítem j del test, que no haya sido administrado todavía. El ítem seleccionado es aquél para el que $I_j(\theta_i)$ tome el mayor valor.

- *Método bayesiano de la máxima precisión esperada.* Fue propuesto por Owen (1975), y selecciona aquella cuestión que minimiza la esperanza de la varianza de la distribución del conocimiento del alumno a posteriori, contribuyendo de esta forma a maximizar la precisión en la estimación. Sea un alumno i con nivel de conocimiento estimado θ_i , que ha respondido previamente a n ítems; y $\mathbf{u}_n = u_1, u_2 \dots u_n$ el vector de respuestas seleccionadas por el alumno en cada ítem administrado. Para calcular cuál es el siguiente ítem que debe ser administrado al alumno, por cada ítem del banco que no haya sido todavía administrado, se calcula la siguiente función:

$$E_u[\sigma^2(\theta_i|\mathbf{u}_n, u_{n+1})] = \sum_{u=0}^1 \sigma^2(\theta_i|\mathbf{u}_n, u_{n+1}) \int P(u_{n+1} = u|\theta_i) p(\theta_i|u_n) d\theta_i \quad (2.32)$$

donde E_u representa la esperanza matemática, y σ^2 la varianza. Siendo u_{n+1} la respuesta que daría el alumno a ese nuevo ítem. $P(u_{n+1} = u|\theta_i)$ representa la CCI del ítem candidato, y $p(\theta_i|u_n)$ la distribución del conocimiento del alumno después de haberle sido administrados los n primeros ítems. En este caso, puede apreciarse que la ecuación 2.32 se ha aplicado a modelos dicotómicos en los que las respuestas posibles son 0 (incorrecta) o 1 (incorrecta).

- *Método basado en el nivel de dificultad.* Propuesto también por Owen (1975), este método selecciona aquel ítem cuya dificultad está más próxima al nivel de conocimiento actual del alumno:

$$\min_{\forall k \in K} |b_k - E(\theta|u_{i1} \dots u_{in})| \quad (2.33)$$

donde K es el conjunto de ítems que no han sido aún administrados, b_k la dificultad del ítem k y $E(\theta|u_{i1} \dots u_{in})$ el nivel de conocimiento estimado, según el criterio EAP después de haber administrado n ítems.

Owen demostró que, para una CCI logística continua, el método bayesiano es equivalente a este método basado en la dificultad, si todos los ítems tienen la misma discriminación. Birnbaum (1968) probó además, que en caso de que el factor de adivinanza sea igual a 0, el ítem que debe seleccionar este criterio, es aquél con la dificultad más cercana al nivel estimado del alumno y con el factor de discriminación más alto. Este método tiene la ventaja de que requiere un menor número de cálculos que el anterior. Su limitación principal radica en que la selección del siguiente ítem que va a ser administrado, se hace en función de la dificultad del ítem, sin tener en cuenta al resto de parámetros que describen su curva característica.

Los dos primeros métodos de selección tienen la desventaja de que, en igualdad de condiciones, van a seleccionar aquellos ítems más discriminantes. Esto puede conllevar una sobreexposición de los ítems con factores de discriminación mayores, por lo que sería necesario aplicar algún método para controlar la exposición de éstos (Revuelta et al., 1998). Otra desventaja está en la falta de mecanismos que aseguren que el contenido de los tests esté bien balanceado. Supongamos que se quiere realizar una versión adaptativa del test para obtener el carné de conducir. Tal y como están definidos los criterios de selección, el algoritmo fácilmente podría limitarse a administrar ítems sobre el tema de "primeros auxiliares", obviando el resto de contenidos. Esta situación es especialmente preocupante cuando los ítems de un cierto tema tienen factores de discriminación más altos que los del resto de temas. Algunas propuestas como la realizada por Kingsbury y Zara (1989), se basan en que el creador del test establezca el porcentajes de ítems que deben administrarse de cada área incluida en el test. Este criterio heurístico se combina con el método de selección adaptativa de ítems utilizado, para asegurar tests balanceados en contenido. Como se verá en capítulos posteriores, el modelo de diagnóstico propuesto en esta tesis incluye un mecanismo (no heurístico) para el balanceo en contenido de tests sobre múltiples conceptos.

Criterios de finalización

El criterio de finalización del algoritmo adaptativo es el encargado de determinar cuándo debe terminar el test. Idealmente, el test debería finalizar cuando la estimación del conocimiento del alumno ha alcanzado un grado de precisión suficiente, es decir, cuando el error en la medida es menor que un cierto umbral predefinido. Además del criterio de la precisión, existen diversos criterios (Olea y Ponsoda, 2001):

- *Criterio de longitud fija:* Según el cual el test finaliza cuando se han administrado un determinado número de ítems. Este método tiene la ventaja de que es de fácil implementación. La limitación que presenta es que dará lugar a examinandos con estimaciones de diferente precisión.
- *Criterio de tiempo límite:* En este caso, el test termina después de que transcurra un tiempo predefinido. Presenta la misma limitación que el criterio anterior.
- Un *procedimiento especial de longitud variable*: Según el cual el test termina si el nivel estimado se aleja de forma significativa de un punto de corte establecido.

En la práctica, los criterios anteriores se utilizan combinados entre sí. De hecho, lo más recomendable es combinar un límite máximo de ítems administrados, con el superar un cierto umbral de precisión en la estimación. El primero de ellos para evitar que el test no sea demasiado largo, y el segundo para asegurar una estimación con un cierto grado de validez.

La aplicación de criterios de longitud variable de ítems, aunque tiene la ventaja de que la finalización del test está motivada porque se ha alcanzado la precisión requerida, presenta la desventaja de que algunos examinandos tienden a sentirse discriminados con respecto a otros. La razón principal es que puede darse la situación de que un alumno, cuyo test tuviera un número menor de ítems, haya obtenido un nivel de conocimiento mayor en la escala que otro que realizó un número mayor de ítems. Para evitar estos problemas, algunos sistemas como CARAT (Segall y Moreno, 1999), emplea el criterio mixto de un máximo de ítems o un error de medida inferior a un cierto umbral.

Durante el proceso iterativo de ejecución del algoritmo adaptativo se van calculando estimaciones temporales del conocimiento del alumno, en función de las respuestas elegidas en cada ítem. Una vez finalizado el test, el valor del nivel de conocimiento inferido después de que todas los ítems del test hayan sido administrados, se convierte en el resultado final del algoritmo adaptativo, y por tanto, en la estimación definitiva del conocimiento del alumno.

En diversas ocasiones, la estimación final del conocimiento del alumno es transformada antes de mostrársele. Por regla general, el nivel obtenido se expresa en la misma escala que si se tratara de un algoritmo convencional de evaluación mediante tests. Esto es, en una escala sobre el total de ítems administrados. Entre los procedimientos que se utilizan para llevar a cabo la transformación, puede destacarse el uso de la denominada *función característica del test* (Lord, 1980). Esta función relaciona el nivel de conocimiento obtenido en el test, con la correspondiente puntuación; y se calcula sumando todas las CCI de los ítems del test. El valor obtenido por el alumno será, de esta forma, la correspondiente al nivel de conocimiento obtenido, según la función característica del test. Otro procedimiento alternativo es la transformación equipercentil de la escala de niveles de conocimiento en la del número total de ítems del test.

2.9.2. Calibración de la CCI

El objetivo de la calibración es inferir la curva característica real que corresponde a cada ítem, según el modelo basado en la TRI elegido para caracterizarla. Como los modelos de la TRI son en general paramétricos, en los que las CCI vienen caracterizadas por un conjunto de parámetros, el problema se reduce a estimar estos últimos. Inicialmente se asume que todos los parámetros del ítem son desconocidos. La única información de la que se dispone a priori son las respuestas de los sujetos (Martínez Arias, 1995). Es por tanto necesario llevar a cabo, previamente a la calibración, un fase de administración de los ítems que se desean calibrar a través de tests no basados en la TRI. A partir de estos datos, se deberá realizar la estimación. Hasta cierto punto, el problema es similar al análisis de regresión, en el que deben estimarse los parámetros de un modelo a partir de respuestas observadas, a través de las denominadas variables predictoras. No obstante, los problemas de estimación en la TRI son más complejos que en el caso de la regresión lineal, ya que en esta última el modelo es lineal y las variables independientes son observables. Por el contrario, los modelos de la TRI son no lineales y la variable regresora θ no es observable. Si θ fuera observable o conocida, el problema de la estimación de los parámetros del ítem se simplificaría considerablemente. Otra complicación añadida es que los modelos no se ajustan siempre perfectamente a los datos.

Ciertos estudios empíricos (Hetter et al., 1994) concluyen que la calibración realizada a partir de tests de lápiz y papel proporcionan resultados comparables a las que se obtienen en la aplicación informatizada de los mismos. Este dato es importante ya que este segundo procedimiento siempre es más costoso (Olea y Ponsoda, 2001).

Anclaje y equiparación

Aunque lo deseable sería administrar a todos los examinandos el banco de ítems completo, se corre el riesgo de potenciar efectos negativos que puedan deteriorar la calidad de las respuestas (fatiga, desmotivación, etc.). Esta limitación no sólo afecta a la calibración de grandes bancos, sino también a los pequeños ya que éstos exigen mayor esfuerzo. Por el contrario, un TAI será eficiente en la medida en que disponga de muchos ítems donde escoger, ya que se atenuarán los factores no controlados que puedan afectar a la estimación (Renom y Doval, 1999).

Para poder obtener las respuestas con las que calibrar los ítems, se suele organizar un procedimiento por el cual éstos se distribuyen en diversos bloques (que serán administrados a través de tests convencional) que comparten algunos ítems en común. Al procedimiento se le denomina *anclaje*, y los ítems comunes se denominan *ítems o test de anclaje*. Cada test se aplica a un conjunto diferente de alumnos. A continuación se lleva a cabo el denominado proceso de *equiparación*. Su objetivo es tener alguna referencia común, que sirva de anclaje en la equivalencia de las métricas realizadas en los distintos tests administrados para llevar a cabo la calibración. Como la función característica de cada ítem es invariante, existirá una relación lineal entre las estimaciones de los parámetros obtenidas para los ítems de anclaje y entre las del parámetro de nivel de conocimiento de los sujetos de anclaje; dado que lo que varía entre ellas es el origen y la unidad de medida. De esta forma, el problema de la equiparación se reduce a determinar el valor de los parámetros que describen esa relación (Barbero, 1996).

Como consecuencia, en general, antes de proceder con la calibración de los ítems, hay que tomar las siguientes decisiones: (1) Determinar el tamaño mínimo muestral que se va utilizar como información de entrada al proceso. (2) Decidir si se va a aplicar un procedimiento de anclaje, y en caso afirmativo, determinar el diseño de anclaje y equiparación. Una vez calibradas las CCI, debe comprobarse el grado de ajuste de las curvas de los ítems al modelo TRI seleccionado, así como otras propiedades psicométricas adicionales.

A continuación se van a describir los métodos de calibración más populares, que se aplican a modelos paramétricos. Posteriormente, y por su relación directa con la técnica de calibración utilizada en esta tesis, se procederá a estudiar la calibración de modelos no paramétricos.

Métodos de calibración de modelos paramétricos

A partir de la información obtenida en forma de sesiones de tests administrados convencionalmente, se procede a calibrar los ítems que ha sido utilizados. Para ello, al comienzo del proceso de calibración, es necesario partir de unos valores iniciales de los parámetros. Éstos son de suma importancia, ya que ellos determinarán la duración del proceso de calibración. La disponibilidad de un peritaje de expertos sobre los valores de los parámetros estimados de los ítems puede ser una pieza clave en el proceso de calibración (Hetter et al., 1994).

En la literatura, se pueden encontrar principalmente, tres aproximaciones para llevar a cabo la calibración de modelos paramétricos. Todos estos procedimientos se hacen inmanejables sin la ayuda de un computador (Santisteban, 1990, cap.12). Estas aproximaciones se basan en la función de máxima verosimilitud (ecuación 2.28). La diferencia entre ellas radica en la forma de conceptualizar la probabilidad de los patrones de respuesta observados (Embretson y Reise, 2000). Todos ellos estiman simultáneamente los parámetros que

caracterizan los ítems, así como el valor del rasgo latente para los alumnos pertenecientes a la muestra utilizada en la calibración:

- *Máxima Verosimilitud Conjunta* (en inglés, *Joint Maximum Likelihood*): Este método (Birnbaum, 1968; Lord, 1980) es un procedimiento iterativo que consta de dos fases. En la primera se estiman los niveles de conocimiento de los sujetos, mientras que en la segunda se estiman los parámetros de los ítems. En la primera iteración de este algoritmo, se inicializan los valores de los parámetros del ítem. A partir de éstos, se estima el nivel de conocimiento de los sujetos de la muestra. A continuación se vuelven a estimar los parámetros de los ítems a partir de la estimación del conocimiento de uno de los alumnos de la muestra. En la segunda iteración, los niveles de conocimiento vuelven a ser estimados con los nuevos valores de los parámetros. Posteriormente, con estos datos, se vuelven a estimar los parámetros de nuevo. El procedimiento continuará hasta que la actualización de los parámetros no sea significativa.

Entre las ventajas de este método destaca su uso generalizado en muchos modelos basados en la TRI, y que se trata de un método computacionalmente eficiente. Entre las desventajas se pueden señalar las siguientes: los valores estimados de los parámetros de los ítems están sesgados, es decir, tienen errores de precisión; y no pueden utilizarse para el proceso de calibración sesiones en las que los alumnos hayan acertado o fallado todas las respuestas.

- *Máxima Verosimilitud Marginal* (en inglés, *Marginal Maximum Likelihood*): En esta técnica de calibración los patrones de respuesta de las sesiones de tests se tratan como la esperanza de una distribución de población. Bock y Aitkin (1981) desarrollaron un algoritmo de esperanza/maximización (EM) para llevar a cabo la estimación. Inicialmente se contabiliza el número de veces que cada patrón de respuesta aparece en la muestra, de tal forma que cada sesión no se tiene en cuenta como tal, sino como una ocurrencia de cierto patrón. Asimismo se considera una distribución normal a priori de los valores que puede tomar el rasgo latente. A esta distribución $P(\theta)$ se le aplica una *cuadratura gaussiana*, es decir, la distribución, cuyo rango de valores original es continuo, se discretiza, dividiendo el rango en un número Q de segmentos iguales. Cada segmento q -ésimo estará representado por un valor entero θ_q . Para llevar a cabo la estimación se utiliza la *probabilidad marginal* de cada patrón de respuesta p que viene definida de la siguiente forma:

$$P(X_p|\beta) = P(X_p) = \sum_q^Q P(X_p|\theta_q, \beta)P(\theta_q) \quad (2.34)$$

donde β es el conjunto de parámetros de los ítems; y $P(X_p|\theta_q, \beta)$ es la función de verosimilitud para el patrón p en el nivel de conocimiento correspondiente al punto de cuadratura q -ésimo. De esta forma, se calcula el valor de la máxima verosimilitud para cada patrón de respuesta:

$$L(\vec{X}) = \prod_p P(X_p|\beta)^{n_p} \quad (2.35)$$

siendo n_p el número de ocurrencias del patrón p . Las ecuaciones de estimación quedan de la siguiente forma:

$$\sum_q x'_{iq} - \sum_q N'_q P(X_{ij} = 1 | \theta_q, \beta) = 0 \quad (2.36)$$

donde x'_{iq} representa el número esperado de sujetos de la muestra con nivel de conocimiento θ_q que han acertado el ítem i , esto es, $x_{ip} = 1$; y N'_q el número esperado de sujetos de la muestra con nivel de conocimiento θ_q . Ambas se calculan de la siguiente forma:

$$n'_{iq} = \sum_p n_p x_{ip} \frac{P(X_p | \theta_q) P(\theta_q)}{P(X_p)} \quad (2.37)$$

$$N'_q = \sum_p n_p \frac{P(X_p | \theta_q) P(\theta_q)}{P(X_p)} \quad (2.38)$$

El procedimiento de estimación se basa en la aplicación del algoritmo de EM, que se compone de dos fases: (a) *Esperanza*: donde se calculan n'_{iq} y N'_q . (b) *Maximización*: a partir de los datos anteriores se calculan los valores de los parámetros de las CCI que maximizan la ecuación 2.36. A continuación, el algoritmo volvería a ejecutar la fase de esperanza hasta que los valores converjan.

- *Máxima Verosimilitud Condicional*: (en inglés, *Conditional Maximum Likelihood*) Sólo es aplicable a modelos 1PL y de gran coste computacional lo que lo hace impracticable en muchos casos.

Tal y como se explica en (Embretson y Reise, 2000), en este procedimiento se parte de la base de que la probabilidad de un determinado patrón de respuesta X_s , dada la puntuación heurística obtenida en el test x_s , y la dificultad del ítem β viene dada por el siguiente ratio:

$$P(X_s | x_s, \beta) = \frac{P(X_s | \theta_s, \beta)}{P(x_s | \theta_s, \beta)} \quad (2.39)$$

donde $P(X_s | \theta_s, \beta)$ es la probabilidad del patrón el nivel de conocimiento θ_s del alumno s , y $P(x_s | \theta_s, \beta)$ la probabilidad de la puntuación del alumno en el test dado su nivel de conocimiento real.

Tómense los parámetros ξ_s como el antilogaritmo del nivel de conocimiento, esto es: $\xi_s = e^{\theta_s}$; y ε_i como el antilogaritmo de la dificultad del ítem, también denominado *facilidad del ítem*, esto es: $\varepsilon_i = e^{-\beta_i}$. Utilizando estos nuevos parámetros, y realizando algunas transformaciones, el numerador y denominador de la ecuación 2.39 pueden expresarse de la siguiente forma:

$$P(X_s | \theta_s, \varepsilon) = \prod_i P_i^{x_{is}} (1 - P_i)^{1-x_{is}} = \frac{\xi_s^{x_s} \prod_i \varepsilon_i^{x_{is}}}{\prod_i (1 + \xi_s^{x_s} \varepsilon_i)} \quad (2.40)$$

$$P(x_s = r | \xi_s, \varepsilon) = \frac{\xi_s^{x_s} \sum_{\sum_i x_{i=r}} \prod_i \varepsilon_i^{x_{is}}}{\prod_i (1 + \xi_s^{x_s} \varepsilon_i)} \quad (2.41)$$

donde x_{is} es la respuesta que el alumno s da al ítem i , P_i la CCI de ese ítem i , y $\sum_i x_{i=r}$ es la suma a lo largo de todos los patrones de respuesta con puntuación r .

Si se dividen las expresiones 2.40 y 2.41, y llamando γ_r a la siguiente expresión $\gamma_r = \sum_{\sum_i x_i=r} \prod_i \varepsilon_i^{x_{is}}$, la ecuación 2.39 se puede formular de la siguiente manera:

$$P(X_s|r, \varepsilon) = \frac{\prod_i \varepsilon_i^{x_{is}}}{\gamma_r} \quad (2.42)$$

El denominador de la expresión anterior es una función simétrica elemental que refleja el aspecto combinatorio de la probabilidad de la puntuación r . Utilizando esta propiedad, la probabilidad de que un ítem concreto i sea correcto, dada la puntuación total en el test, puede expresarse de la siguiente forma:

$$P(X_s = 1|r, \varepsilon) = \varepsilon_i \frac{\gamma_{r-1}^{(i)}}{\gamma_r} \quad (2.43)$$

donde $\gamma_{r-1}^{(i)}$ es la función simétrica elemental de $r - 1$ argumentos, en la que el valor de facilidad del ítem se omite.

A partir de lo anterior, la probabilidad de un determinado patrón de respuesta se podría expresar de la siguiente forma:

$$P(\mathbf{X}|r, \varepsilon) = \prod_s \prod_i \frac{\varepsilon_i^{x_{is}}}{\gamma_r} \quad (2.44)$$

Si se aplican logaritmos y se expresa en función de la dificultad, la ecuación anterior quedaría:

$$\ln P(\mathbf{X}|r, \beta) = - \sum_i x_i \beta_i - \sum_s \ln \gamma_r s \quad (2.45)$$

Así, el sistema de ecuaciones que se deberá resolver para llevar a cabo la calibración de los I ítems queda de la siguiente forma:

$$\begin{aligned} 0 &= \sum_s x_{1s} - \sum_s e^{-\beta_1} \frac{\gamma_{r-1}^{(1)}}{\gamma_r} \\ 0 &= \sum_s x_{2s} - \sum_s e^{-\beta_2} \frac{\gamma_{r-1}^{(2)}}{\gamma_r} \\ &\dots \\ 0 &= \sum_s x_{Is} - \sum_s e^{-\beta_I} \frac{\gamma_{r-1}^{(I)}}{\gamma_r} \end{aligned} \quad (2.46)$$

Para más información sobre este método de calibración, véase (Wainer et al., 1980).

El cálculo en estos métodos de calibración se basa en la aplicación de métodos numéricicos de aproximación como el de Newton-Raphson, que es un proceso de búsqueda iterativo en el cual las estimaciones de los parámetros son sucesivamente mejoradas. En este procedimiento es necesario establecer a priori: (a) las condiciones que deben cumplirse para que la estimación se considere satisfactoria; (b) cómo mejorar las estimaciones obtenidas; y (c) un criterio para determinar cuándo debe finalizar el proceso de calibración.

Los métodos anteriores de calibración se utilizan para modelos que se basan en la función 1PL, 2PL o en la 3PL. Estos modelos sólo deben utilizarse en casos en los que se dispone de gran cantidad de ítems y de sesiones de test. La *calibración bayesiana* surge de la necesidad

de buscar métodos de estimación para funciones 3PL, en los casos en los que el volumen de información del que se dispone es de tamaño medio o incluso pequeño (Mislevy, 1986). En la calibración bayesiana, se determinan probabilidades a priori para los parámetros. Existen versiones bayesianas de los métodos de máxima verosimilitud conjunta y marginal.

En el ámbito de los TAI, el procedimiento ideal para llevar a cabo la estimación de parámetros de las CCI se compone de dos fases (Glas, 2000): La primera es la *calibración inicial*, donde se realizan tests con ítems que no han sido calibrados. Esta fase suele reemplazarse por un test de papel y lápiz, aunque lo más adecuado es utilizar el propio entorno que sustentará los tests calibrados, ya que esto asegura una calibración más precisa. En la segunda fase, la *calibración en línea*, el objetivo es, partiendo de ítems ya calibrados, aplicar el procedimiento con ítems nuevos. De esta forma, el alumno realiza un test en el que hay ítems que están calibrados y otros que no lo están.

Para obtener más información sobre los procedimientos de calibración paramétricos véase (Wainer y Mislevy, 1990; Santisteban, 1990; Glas, 2000; Embretson y Reise, 2000).

Métodos de calibración de modelos no paramétricos: El suavizado núcleo

Para la calibración en modelos no paramétricos, se suelen adaptar métodos de regresión no paramétricos (Habing, 2001). Muchos de estas técnicas se basan en el cómputo de un promedio local, ya que el cálculo de la función de regresión $F(X)$ es equivalente al cálculo de su esperanza matemática para cada valor $E(Y|X = x)$. Por este motivo, es razonable asumir que los valores de $F(X)$ pueden obtenerse tomando una media ponderada de la variable de respuesta Y sobre aquellos alumnos cuyo nivel de conocimiento está más cercano a x . Según el método de suavizado que se aplique, esa ponderación se obtendrá de una manera u otra (Douglas y Cohen, 2001).

Siguiendo esta línea, una de las técnicas más utilizadas, principalmente por su simplicidad, es el *suavizado núcleo* (en inglés, *kernel smoothing*) (Härdle, 1992; Wand y Jones, 1995; Simonoff, 1996). Ya en 1857, Engel la utilizó para construir una curva denominada regresograma. A pesar de ser técnicas estadísticas antiguas, según Härdle (1992), la parte fundamental de su teoría y sus métodos han sido desarrollados desde mediados de la década de los ochenta. El auge tardío de este conjunto de técnicas se ha debido a que se preferían las aproximaciones paramétricas por ser computacionalmente más simples. La razones que han propiciado este auge son la falta de flexibilidad de los modelos paramétricos en lo que a análisis de datos se refiere, y el desarrollo del software que permite realizar estimaciones no paramétricas.

La idea principal que subyace en las técnicas de suavizado es que, dada una función $m(x)$ y un conjunto de observaciones X , aquéllas que estén cerca de x_i contendrán información sobre el valor de m en x_i . Por tanto, para hacer una estimación del valor $m(x_i)$ es posible utilizar una especie de promedio local de los datos cercanos a x_i (Eubank, 1988). Para llevar a cabo el promedio local, durante el cálculo del valor estimado, se ponderan los valores de sus vecinos multiplicándolos por un conjunto de pesos. Esta secuencia de pesos se suele representar mediante una función de densidad con un parámetro de escala que se encarga de ajustar el tamaño y la forma de los pesos cercanos a x . Suele ser bastante común referirse a esta función como la *función núcleo K suavizado núcleo!función núcleo*.

Entre las características de la función núcleo, se pueden destacar que es continua, que no es infinita (está limitada), tiene su máximo valor en $u = 0$, y es simétrica con respecto al eje de ordenadas $x = 0$. Los valores $K(u)$ son siempre mayores o iguales a cero, y tiende a cero

conforme u se separa de cero en cualquier dirección, es decir, $\lim_{u \rightarrow +\infty} = \lim_{u \rightarrow -\infty} = 0$. Por último, la función núcleo es simétrica real, y su integral es igual a 1:

$$\int K(u)d(u) = 1$$

Ramsay (1991) fue el primero en popularizar el uso de técnicas de suavizado para la calibración dentro de la TRI, proporcionando una método de calibración relativamente sencillo de implementar (Junker y Sijtsma, 2001b). El trabajo de Ramsay no fue más que el punto de partida para muchos otros investigadores (Drasgow et al., 1992; Samejima, 1998; Douglas y Cohen, 2001; Ferrando, 2004), que en estos últimos años han aplicado este tipo técnicas. Asimismo el propio Ramsay (2000) es el responsable de una herramienta software de libre disposición, TESTGRAF, para la calibración no paramétrica de ítems aplicando suavizado núcleo.

El procedimiento de calibración aplicando suavizado es un algoritmo que, al igual que los métodos convencionales, utiliza como información de entrada sesiones de tests. Éstas, realizadas por una muestra de alumnos de forma no adaptativa y sin aplicar la TRI para evaluar, se basan en la administración de los ítems que se desea calibrar, y se evalúan aplicando algún criterio clásico.

El algoritmo de calibración basado en el suavizado núcleo permite calibrar no sólo ítems dicotómicos, sino también los de opción múltiple polítómicos. Dado que el modelo de respuesta que se propone en esta tesis es polítómico, en esta sección, se abordará la calibración de ítems de este tipo, utilizando el suavizado núcleo. Nótese que para obtener la versión dicotómica de este algoritmo bastará con considerar que los ítems sólo tienen dos respuestas, esto es, correcta o incorrecta.

Para llevar a cabo la calibración, esta técnica discretiza el dominio de las curvas características (el nivel de conocimiento). Más concretamente, TESTGRAF establece 51 puntos de evaluación θ_q ($Q = 51$), resultantes de discretizar el rango entre $-2,5$ y $2,5$, tomando un punto ($q = 1, \dots, Q$) por cada décima. Considerando esta discretización, las fases de las que consta este algoritmo son las siguientes (Ramsay, 2000):

1. *Clasificación:* Se estima la puntuación T_a del alumno a -ésimo en el test. En su propuesta original, Ramsay sugiere inicialmente utilizar el porcentaje de ítems correctamente respondidos, aunque posteriormente (Ramsay, 1991) apunta que ese método de evaluación puede generar resultados muy sesgados en tests de pocos ítems aplicados a una población con un tamaño considerable de examinandos. Asimismo indica que ese criterio no tiene en cuenta el ítem en sí, puesto que, en general, suele haber ítems con más calidad que otros. Por este motivo, sugiere otro estadístico para calcular la evaluación del alumno en el test convencional:

Sea el alumno a , para cada ítem i y para la respuesta j de ese ítem:

$$T_a = \sum_{i=1}^I \sum_{j=1}^J Y_{ij}^a W_{ij} \quad (2.47)$$

donde Y_{ij}^a es igual a uno cuando el sujeto a ha seleccionado la opción j del ítem i , y cero en otro caso. W_{ij} se define de la siguiente forma:

$$W_{ij} = \text{logit}(P_{ij}^{(75)}) - \text{logit}(P_{ij}^{(25)}) \quad (2.48)$$

siendo $\text{logit}(p) = \ln(\frac{p}{1-p})$, y $P_{ij}^{(75)}$ y $P_{ij}^{(25)}$ la proporción de examinandos que están por encima y por debajo del 25 % de la distribución de porcentajes de ítems acertados, respectivamente.

Este estadístico tiene la ventaja de que hace uso de la información en el caso de que el alumno seleccione una respuesta incorrecta. De igual forma, aquellos ítems que no discriminan lo suficiente entre niveles de conocimiento bajos y altos tienen poca relevancia en la ponderación.

2. *Enumeración:* Se reemplaza los T_a por los cuantiles de la distribución normal estándar, que son los valores que dividen el área bajo la función de densidad normal estándar en $N+1$ áreas de tamaño $\frac{1}{N+1}$. Estos valores se asumirán como los niveles de conocimiento θ_a de los N alumnos de la muestra.
3. *Ordenación:* Los patrones de respuesta de los alumnos $(X_{a1}, X_{a2}, \dots, X_{an})$ se ordenan según la clasificación anterior. Es decir, el a -ésimo patrón de respuesta según la ordenación $(X_{(a)1}, X_{(a)2}, \dots, X_{(a)n})$, es el a -ésimo alumno según su θ_a .
4. *Suavizado:* Para la m -ésima opción del ítem i -ésimo, se estima su CCR_{im} aplicando suavizado a la relación existente entre el vector de tamaño N con los valores de la variable binaria y_{ima} que indica si el alumno a ha seleccionado la opción m en el ítem i , y el vector con los niveles de conocimiento de cada sujeto $\theta_1, \dots, \theta_N$. El cálculo de la curva característica se lleva a cabo aplicando la siguiente fórmula:

$$CCR_{im}(\theta_q) = \sum_{a=1}^N w_{aq} y_{ima} \quad (2.49)$$

donde cada w_{aq} se calcula de la siguiente forma:

$$w_{aq} = \frac{K(\frac{\theta_a - \theta_q}{h})}{\sum_{b=1}^N K(\frac{\theta_b - \theta_q}{h})} \quad (2.50)$$

siendo K la *función núcleo* (en inglés, *kernel function*), y h el *parámetro de suavizado o ancho de banda* (en inglés, *smoothing parameter* o *bandwith*). Este último controla el tamaño de la medida del desplazamiento (Ramsay, 2000). Es decir, controla el radio de influencia de los valores vecinos sobre el que se desea estimar. Analíticamente, cuanto menor sea el valor de h , menos se reducirá la diferencia $\theta_a - \theta_j$, y por tanto, mayor será el radio de influencia de los valores contiguos sobre el estimado.

Ramsay (1991) tras varios experimentos determina que el valor más adecuado para h se calcula utilizando el siguiente heurístico: $h = N^{-1/5}$ siendo N el tamaño de la muestra poblacional utilizada en el proceso de calibración. Posteriormente, el propio Ramsay (2000) modifica ligeramente este heurístico, quedando de la siguiente forma: $h = 1,1N^{-1/5}$, siendo este valor el que utiliza en su programa de calibración TESTGRAF.

Ramsay sugiere la posibilidad de repetir el procedimiento anterior, aplicando un proceso de refinamiento, de forma que se vuelvan a estimar las CCR, hasta que la estimación se estabilice; es decir, hasta que la variación entre iteraciones no sea significativa. La idea que propone es repetir todo el proceso, pero sustituyendo las puntuaciones en el test convencional por el nivel de conocimiento estimado de cada alumno, aplicando la máxima verosimilitud

(ecuación 2.28) utilizando las CCR recién estimadas. Como se verá en la sección 4.7, en el marco de esta tesis se propone una modificación de este algoritmo que mejora sus resultados.

En el suavizado núcleo, como función núcleo se puede utilizar cualquiera que cumpla las condiciones anteriormente mencionadas. Las tres funciones más comúnmente utilizadas (Ramsay, 1991) se enumeran a continuación:

- *Función uniforme:*

$$K(u) = \begin{cases} 1 & \text{si } -1 \leq u \leq 1 \\ 0 & \text{en otro caso} \end{cases} \quad (2.51)$$

- *Función cuadrática o Epanechnikov:*

$$K(u) = \begin{cases} 1 - u^2 & \text{si } -1 \leq u \leq 1 \\ 0 & \text{en otro caso} \end{cases} \quad (2.52)$$

- *Función gaussiana:*

$$K(u) = e^{\frac{-u^2}{2}} \quad (2.53)$$

2.9.3. TAI con modelos politómicos

Mecanismos de selección de ítems

Los métodos de selección más utilizados en TAI politómicos son los basados en la información suministrada por el ítem sobre el nivel de conocimiento estimado (Hontangas et al., 2000). Dentro de esta línea, existen diversos criterios dependiendo de la función de información que se utilice. Pueden utilizarse la función de información del ítem, la de las alternativas o bien la función de información en un intervalo alrededor del nivel de conocimiento estimado.

El caso más general es el que utilizan Dodd et al. (1995), que se basa en aplicar la siguiente fórmula:

$$I_j(\theta) = \sum_{x=0}^m m \frac{P'_{jx}(\theta)^2}{P_{jx}(\theta)} \quad (2.54)$$

donde $I_j(\theta)$ es la *función de información* del ítem j , que tendrá $m + 1$ posibles respuestas, y la curva característica de la respuesta x es $P_{jx}(\theta)$.

Otra criterio alternativo al anterior, se basa en el cálculo de la denominada *función de información de la alternativa x* , $I_{jx}(\theta)$, que se define de la siguiente forma:

$$I_{jx}(\theta) = -\frac{\delta^2}{\delta \theta^2} \ln P_{jx}(\theta) \quad (2.55)$$

A partir de ésta, se define la denominada *contribución de la alternativa a la información del ítem* $I_{jx}^*(\theta)$ definida como:

$$I_{jx}^*(\theta) = I_{jx}(\theta) P_{jx}(\theta) \quad (2.56)$$

Por último, se define la *función de información del ítem j* basada en la contribución de las alternativas , $I_j^*(\theta)$, de la siguiente forma:

$$I_j^*(\theta) = E[I_{jx}(\theta)] = \sum_{k=0}^m I_{jx}^*(\theta) = \sum_{x=0}^m I_{jx}(\theta)P_{jx}(\theta) \quad (2.57)$$

Sea cual sea el criterio basado en la función de información utilizado, el ítem elegido es siempre aquél cuyo valor de función de información, para el nivel de conocimiento estimado, es mayor. Según De Ayala (1992), cuando se utiliza $I_j^*(\theta)$ frente a $I_j(\theta)$ se obtiene una pequeña reducción en el número de ítems del test. Otros criterios de selección aplicados a ítems polítómicos son el criterio de máxima dificultad, y la mínima entropía.

Mecanismos de estimación del conocimiento

En cuanto a la estimación del nivel de conocimiento del alumno en TAI con modelos de respuesta polítómicos, se utilizan métodos análogos a los utilizados en los dicotómicos. La única variación sobre éstos está en que, en vez de utilizar la CCI o su inversa, en función de si la respuesta es correcta o no, se utiliza la CCR de la respuesta seleccionada por el alumno. Por consiguiente, la función de máxima verosimilitud para el caso polítómico vendría descrita de la siguiente forma:

$$P(\theta|u) = L(u|\theta) = \prod_{i=1}^n P_{ik_i}(\theta) \quad (2.58)$$

donde k_i es la respuesta seleccionada por el alumno en el ítem i , y u el patrón de respuestas $u = u_1, \dots, u_n$.

Criterios de finalización

Por último, los criterios de finalización para tests polítómicos también son análogos a los dicotómicos. Igualmente, el test suele finalizar cuando la precisión de la estimación del conocimiento del alumno es lo suficientemente alta, o cuando se ha sobrepasado el número máximo de ítems fijados por el profesor.

2.10. Los sistemas de tests comerciales

En el mercado existen gran cantidad de herramientas comerciales de evaluación basadas en tests. La mayoría de ellas implementan tests con criterios de evaluación convencionales, tales como el porcentaje de ítems respondidos correctamente, o la suma de puntuaciones de respuestas correctas. Por el contrario, existen algunas pocas que implementan TAI.

En esta sección se procederá a estudiar algunas de estas herramientas disponibles para la construcción y administración de tests. Desafortunadamente, por su carácter comercial, los desarrolladores de estos sistemas no suministran mucha información sobre sus características.

2.10.1. Sistemas no adaptativos

Intralearn En este sistema (Intralearn Software Corp., 2003) cada ítem definido por el profesor está asignado a un único test. Asimismo los ítems pueden asociarse a diversos subtemas dentro de una jerarquía curricular. El sistema almacena modelos del alumno, los cuales son accesibles en cualquier momento. Intralearn define dos estrategias de evaluación: *normalizada*, en la que se obtiene el porcentaje de preguntas acertadas sobre el total; y la *media ponderada*, donde cada ítem tiene asociado una puntuación en caso de que sea correctamente respondido. En cada test, los profesores deben indicar la puntuación o el porcentaje de preguntas acertadas necesarias para considerar que el examinando ha aprobado. El sistema incluye además la posibilidad de incluir refuerzos junto con la corrección del ítem.

WebCT En esta herramienta (WebCT Inc., 2003) el material educativo está estructurado en conceptos y subconceptos. Cada uno de ellos tendrá asociado un banco de ítems. Cada test puede estar formado por ítems de diferentes conceptos, pudiendo cada ítem formar parte de diferentes tests. En cuanto a los tests, su disponibilidad puede limitarse a un periodo de tiempo, y se pueden definir los siguientes tipos: cuestionarios, tests con preguntas de opción múltiple, etc. Después de que el examinando responda a cada ítem, junto con la corrección se le suministra un refuerzo. El sistema facilita los siguientes tipos de ítems: de opción múltiple, de respuesta múltiple, de correspondencia, de respuesta corta y de redacción. A excepción del último tipo, el resto se evalúan automáticamente en línea durante la realización del test. Los ítems de redacción, por el contrario, tienen que ser evaluados a posteriori por el profesor de forma manual. Al igual que el sistema anterior, WebCT ofrece dos mecanismos de evaluación: porcentaje de ítems respondidos correctamente sobre el total de los que han sido administrados, y puntuación obtenida en ítems puntuados. Además, esta herramienta incluye la posibilidad de penalizar con puntuaciones negativas los ítems respondidos incorrectamente. Asimismo ofrece aplicaciones que muestran estadísticas sobre los datos, que facilitan el análisis de los resultados, y otras opciones para la generación de informes.

QuestionMark Este sistema (QuestionMark Corp., 2003) trae consigo una herramienta de autor fuera de línea para la creación de ítems. Éstos están estructurados en categorías, y se pueden incluir los siguientes tipos: de opción múltiple, de respuesta múltiple, de respuesta corta, de correspondencia, verdadero/falso, de redacción, y por último, ítems de respuesta sobre figuras. Se permite la inclusión de un refuerzo junto con la corrección del ítem. En cuanto a la selección de ítems, pueden definirse tests con un cierto grado de adaptabilidad. El profesor puede crear enlaces a diferentes partes del test en función de las respuestas dadas por el alumno. Asimismo, se permite la selección aleatoria de los ítems de un test, y restringir el tiempo máximo disponible por los alumnos para completar el test. También se pueden crear grupos de usuarios para limitar el acceso a los tests. Por último, el sistema ofrece herramientas para la gestión de grupos de usuarios y para la generación de informes sobre los resultados de los examinandos en los tests.

TopClass En esta herramienta (WBT Systems, 2003) los tests se crean a partir del banco de ítems. El sistema maneja los siguientes tipos de ítems: opción múltiple, de completar, respuesta múltiple, verdadero/falso, de respuesta sobre figura y de correspondencia. Es posible definir un único testlet con todas los ítems del test. En la definición de cada ítem,

el profesor puede restringir la posibilidad de que los alumnos dejen la respuesta en blanco, e incluir refuerzos. Asimismo, existe la posibilidad de que los alumnos puedan dejar un test durante su realización, y volverlo a retomar en el mismo estado. Se incluye el criterio de selección aleatoria de ítems, en el cual el profesor debe indicar el porcentaje de ítems de cada área de contenido del test que deben suministrarse. La evaluación se basa en el porcentaje de ítems correctamente respondidos o en el uso de ítems con puntuación. Por último, ofrece una herramienta de consulta de las sesiones de tests realizadas.

I-assess En este sistema (EQL International Ltd., 2003) los ítems se estructuran en carpetas. Los tests se configuran, o bien seleccionando directamente los ítems que van a formar parte del test, o bien seleccionando una carpeta, quedando en este caso, todos sus ítems automáticamente elegidos para formar parte del test. El sistema permite incluir los siguientes tipos: de opción múltiple, de correspondencia, de respuesta corta numérica, de redacción y de respuesta sobre figuras. Además de los anteriores, pueden incluirse ítems por partes e ítems generativos. Los primeros son testlets, y pueden añadirse a un test junto con los demás tipos. Los ítems generativos son plantillas que generan ítems dinámicamente, y que se basan en la inclusión de variables cuyos valores se calculan aleatoriamente. En *I-assess*, también es posible mostrar un refuerzo junto con la corrección del ítem. Asimismo, se pueden crear tests en los que el criterio de selección de ítems sea aleatorio, así como restringir el tiempo total que tienen los alumnos para completarlo. Por último, el sistema ofrece una herramienta para la generación de informes sobre los resultados de los alumnos en los tests.

Webassessor Es una sistema (Drake Kryterion, 2004) con una interfaz web que dispone de una herramienta de autor para la construcción de tests, y otra para su administración. Los tests se construyen con ítems de opción múltiple, de verdadero/falso, de respuesta múltiple, de respuesta corta y de redacción. Los criterios de evaluación que ofrece este sistema son el porcentaje de ítems respondidos correctamente, o bien la suma de los puntos asignados a las respuestas dadas por el alumno. También se pueden definir tests temporizados o tests con selección aleatoria de ítems. El sistema incorpora un módulo de administración que permite gestionar alumnos, generar informes, etc.

C-Quest Es una herramienta (Assessment Systems Corporation, 2004a) para la construcción y administración de tests vía web. Los ítems que permite son de opción múltiple, de verdadero/falso, de respuesta corta y de redacción. Los tests pueden temporizarse, y los ítems permiten añadir ayudas y refuerzos para las respuestas incorrectas. Ofrece asimismo una herramienta de administración para poder analizar las sesiones realizadas por los alumnos.

2.10.2. Sistemas basados en TAI

Las herramientas vistas hasta ahora no implementan TAI. En el mercado existen pocas aplicaciones comerciales de evaluación mediante tests basadas en TAI. En este apartado se incluirán algunas de las que implementan por lo menos algunas de las fases de un TAI.

MicroCAT y FastTEST MicroCAT (Assessment Systems Corporation, 2004c), creada a principio de los años 80, fue la primera herramienta software para la construcción y

administración de TAI en PC. Igualmente permitía administrar tests convencionales y ramificados. Con la llegada de los sistemas operativos sobre Windows, se introdujeron diversas mejoras sobre MicroCAT que dieron lugar a FastTEST (Assessment Systems Corporation, 2004b). Este sistema, incorpora las características propias de MicroCAT, añadiendo además otras nuevas. Se estructura en los denominados *espacios de trabajo* que incluyen bancos de ítems y tests. Cada profesor tiene la posibilidad de crear tantos espacios de trabajo como desee. Los ítems de un banco se estructuran en tantas categorías como se desee, y éstas a su vez pueden dividirse en otras subcategorías y así sucesivamente. La herramienta de edición de ítems incluye un procesador de textos con una utilidad para añadir imágenes, así como diccionarios en diversos idiomas, entre ellos el español. Los ítems que se incluyen en un test pueden ser de opción múltiple, de verdadero/falso, de respuesta múltiple y de respuesta corta, entre otros, ofreciéndose la posibilidad de agruparlos en testlets. Para cada ítem, se estudia su CCI, así como el porcentaje de alumnos que lo han respondido correctamente. Los modelos basados en la TRI que utiliza son los dicotómicos: 1PL, 2PL y 3PL. En cuanto a los tests, pueden definirse a partir de un conjunto de ítems seleccionados de forma aleatoria de entre los pertenecientes a una categoría, aleatoriamente de una o más categorías, o todos los ítems de uno o más bancos. Los criterios de selección que ofrece FastTEST son, en tests convencionales: todas las preguntas se muestran en un orden preestablecido común para todos los examinandos, generación aleatoria previa a la administración del test, o bien generación aleatoria diferente para cada alumno; para los TAI: tests de longitud fija, tests de clasificación o tests en los que la finalización se decide en función de la precisión de la estimación del conocimiento del alumno.

TerraNova CAT TerraNova CAT (*California Achievement Tests*) (CTB McGraw-Hill, 2004) es un sistema *ad hoc* para la construcción y administración de tests temporizados para alumnos de colegios norteamericanos. El sistema incluye tests dicotómicos cuyos ítems han sido modelados con la función 3PL, y tests polítómicos con ítems modelados según el modelo de crédito parcial de dos parámetros. Los tests de este sistema fueron calibrados durante 1998, a partir de una muestra poblacional de más de cien mil alumnos, y utilizando técnicas de anclaje. Para llevar a cabo esta tarea se implementó una herramienta software propia llamada PARDUX (Burket, 1991), que permite estimar simultáneamente las curvas características de los ítems dicotómicos y de los polítómicos, y que utiliza el método de calibración de máxima verosimilitud marginal. Según los autores, las calibraciones realizadas con PARDUX son iguales o incluso mejores que las realizadas con otras aplicaciones. Por último, el sistema incluye una herramienta de autor, ITEMSSYS (Burket, 1988), para que los profesores puedan seleccionar qué ítems deben administrarse, que muestra información sobre cada ítem, y que permite decidir cuántos ítems de cada área de contenido van a incluirse en el banco de ítems del test. En cuanto a los criterios de evaluación utilizados, están el criterio del porcentaje de respuestas acertadas y la evaluación basada en la TRI. Los tests son capaces de evaluar diversas áreas de contenido simultáneamente, proporcionando al final una calificación global y una por cada área.

CATGlobal Es quizás la herramienta (Promissor, 2003) para la construcción de TAI más popular, aunque la información que sobre su funcionamiento ofrece su web es bastante pobre: Implementa un modelo de TRI clásico dicotómico, que sólo permite evaluar un concepto en cada test.

2.10.3. Conclusiones

Esta sección ha estado dedicada a los sistemas de tests comerciales. Las aplicaciones descritas son herramientas versátiles con atractivas interfaces. La mayoría de ellas no permiten construir tests con sistemas de evaluación o selección de ítems bien fundamentados como los TAI. Por otro lado, aquéllos que implementan TAI parecen poco adecuados para su uso en la enseñanza, puesto que, en principio, sólo permiten medir el nivel de conocimiento de forma global en todo el test. Aunque los autores de TerraNova CAT mencionan la posibilidad de administrar tests de contenido balanceado, y que como resultado proporcionan una evaluación granular, no queda muy claro si ésta se limita únicamente a los tests convencionales.

También se observa que la construcción de los bancos de ítems suele ser poco flexible. Los ítem sólo pueden estructurarse en un número limitado de categorías, y la calificación final del test suele hacer referencia a la asignatura completa, sin dar información sobre el conocimiento del alumno en cada categoría. Asimismo, estos sistemas presentan una arquitectura monolítica, lo que dificulta enormemente su integración en aplicaciones de enseñanza. A pesar de estas desventajas aparentes, cuentan con un gran número de adeptos. La razón principal está en que éstos prefieren sacrificar el uso de técnicas de evaluación con una base teórica por interfaces de más alto nivel, justo lo contrario que sucede en la mayoría de STI (Weber y Brusilovsky, 2001). Además, tienen la garantía de que ofrecen un servicio de mantenimiento a los usuarios.

La tabla 2.1 es un resumen de las características de los sistemas comerciales que ha sido analizados en esta sección. Cada fila se corresponde con un sistema. En las columnas se representa cada característica estudiada. Las celdas en blanco indican que el aspecto correspondiente no está especificado en la bibliografía.

2.11. Discusión y conclusiones generales del capítulo

En este capítulo se han presentado los tests como un mecanismo de evaluación en dominios declarativos. Se caracterizan porque pueden utilizarse prácticamente en cualquier dominio de este tipo, siendo por tanto, un mecanismo de evaluación genérico, fácilmente adaptable a cada disciplina. Asimismo, el uso de tests tiene la ventaja de que, en principio, los costes de implementación son relativamente pequeños.

Se ha puesto de manifiesto la existencia de dos teorías que modelan la relación entre el comportamiento observado del alumno durante el test y su nivel de conocimiento. Por una parte, la TCT parece poco adecuada si se quiere llevar a cabo una evaluación del alumno rigurosa e independiente de las características de cada test. Sin embargo, la TRI, y su aplicación más directa, esto es, los TAI, tienen todas aquellas peculiaridades que los convierten en instrumento de inferencia del conocimiento del alumno bien fundamentada y con el valor añadido de que los resultados que se obtienen son independientes del test utilizado.

Según Muñiz y Hambleton (1999), los TAI son el resultado de la simbiosis entre los avances informáticos y las aportaciones de los modelos psicométricos de la TRI. La mayor precisión que aportan los tests adaptativos frente a otros sistemas de evaluación, los convierte en un mecanismo de evaluación muy útil dentro de los sistemas educativos, especialmente en aquellos sistemas que incorporan IA (los STI), tal y como se pondrá de manifiesto en el siguiente capítulo. Según Gouli et al. (2001), en un sistema tutor inteligente, la precisión

	<i>Tipos ítems</i>	<i>Criterio evaluación</i>	<i>Criterio de selección</i>	<i>Criterio finalización</i>	<i>¿Permite estructuración curricular?</i>	<i>¿Permite incluir refuerzos?</i>
Intralearn		porcentaje, por puntos	no adaptativo	número de ítems	Sí	Sí
WebCT	OM, RM, correspondencia, RC, de redacción		ramificado	número de ítems, tiempo límite		Sí
TopClass	V/F, OM, RM, RC múltiple, de figuras, de correspondencia, testlets	porcentaje, por puntos	no adaptativo		ítems estructurados en áreas de contenido	Sí
I-assess	OM, correspondencia, RC numérica, de redacción, de figuras, generativos, testlets		no adaptativo	número de ítems, tiempo límite	ítems estructurados en carpetas	Sí
Webassessor	V/F, OM, RM, RC, de redacción	porcentaje, por puntos	no adaptativo	número de ítems, tiempo límite		
C-Quest	V/F, OM, RC, de redacción		no adaptativo	número de ítems, tiempo límite		Sí
MicroCAT/ FastTEST	RC, testlets	por puntos, basados en TRI dicotómica	TRI y ramificado	número de ítems, tiempo límite, precisión	ítems estructurados en categorías	
TerraNova CAT		basados en TRI dicotómica, crédito parcial	TRI		ítems estructurados en áreas de contenido	
CATGlobal		basados en TRI dicotómica	TRI			

Tabla 2.1: Características generales de los sistemas de tests comerciales.

Leyenda: Las siglas incluidas en la primera columna tienen el siguiente significado: V/F (ítems verdadero/falso), OM (ítems de opción múltiple), RM (ítems de respuesta múltiple) y RC (ítems de respuesta corta).

en las estimaciones del conocimiento del alumno es un aspecto crítico de vital importancia para la eficiencia del sistema, puesto que estas estimaciones se utilizan para llevar a cabo la adaptación, esto es, para determinar el siguiente paso que debe dar el alumno en su proceso de instrucción.

Asimismo, se han estudiado los principales modelos de evaluación basados en la TRI que hacen un tratamiento politómico de la respuesta. Éstos requieren un número menor de ítems para llevar a cabo la estimación del conocimiento del alumno, que incluso es más precisa que utilizando los clásicos modelos dicotómicos. Sin embargo, los politómicos, aunque inicialmente fueron concebidos para ser utilizados en ítems de opción múltiple con respuestas no ordenadas, han sido aplicados en entornos en los que éstas están ordenadas (por ejemplo, escalas de tipo Likert), y principalmente en tests de personalidad, más que en TAI. El principal problema está en que, si ya de por sí es costosa la calibración de ítems dicotómicos, el uso de politómicos multiplica ese coste, ya que por cada uno de ellos es necesario inferir más de una curva característica.

En conclusión, los TAI representan un mecanismo de evaluación que en principio puede

ser adecuado para el diagnóstico del alumno en STI, como ponen de manifiesto Kingsbury y Houser (1993). El principal problema es que los TAI, tal y como se han presentado en este capítulo presentan diversos inconvenientes, algunos de los cuales se enumeran a continuación:

- A través de un TAI sólo se puede obtener como resultado una única estimación del conocimiento del alumno. Esto supone que si se quieren evaluar diversos conceptos de una asignatura es necesario realizar tests diferentes.
- Asimismo, cuando se evalúan tests de múltiples conceptos no se puede asegurar que los ítems que se administran sean de contenido balanceado. Esto implica que el nivel de conocimiento que se infiere no refleja el conocimiento global del alumno en todos esos conceptos.
- Por otra parte, un TAI requiere un banco con un gran número de ítems. Aunque los modelos politómicos reducen este requisito notablemente y además su utilización implica una reducción considerable en el número de ítems de los tests, tienen otras desventajas que hay que sumar a las anteriores, como el hecho de que se aplican a tests de personalidad, o únicamente incluyen ítems de opción múltiple.
- Por último, los TAI requieren un proceso preliminar de calibración de los ítems, antes de poder utilizarse. Para ello, es necesario disponer de una población con muchos alumnos que realicen un test con esos ítems de forma convencional. La calibración en modelos politómicos es si cabe más costosa que en los dicotómicos, puesto que la muestra inicial de examinandos debe ser aún mayor.

Con respecto al último de los problemas, el suavizado núcleo, método de calibración presentado para modelos no paramétricos, es una técnica que requiere un número menor de información para inferir las curvas características de los ítems, y además es computacionalmente menos costosa.

En cuanto a los sistemas de tests comerciales presentados, a pesar de que poseen atractivas interfaces, la mayoría están basados en la TCT, y los que lo están en la TRI ofrecen un conjunto reducido de ítems. En general, los modelos de respuesta que utilizan suelen ser los modelos dicotómicos que usan funciones logísticas (1PL, 2PL o 3PL). Asimismo, no parecen facilitar ninguna utilidad que permita su integración en otro tipo de sistemas, ni permiten una estructuración jerárquica de los ítems en conceptos y subconceptos.

Capítulo 3

El diagnóstico en los Sistemas Tutores Inteligentes

*La naturaleza hace que los hombres
nos parezcamos unos a otros y nos juntemos;
la educación hace que seamos diferentes y nos alejemos.*
Confucio

En el capítulo anterior se han presentado los TAI como un mecanismo de evaluación personalizada, que se caracteriza por su generalidad, ya que pueden ser aplicados para la evaluación del conocimiento declarativo del alumno en gran cantidad de ámbitos. Además, este tipo de tests tienen la ventaja de que se basan en una teoría psicométrica bien fundamentada, la TRI. Por este motivo, son capaces de asegurar, además de una evaluación personalizada, un proceso de inferencia del estado de conocimiento del alumno preciso y fiable.

En este capítulo, se estudian todos los paradigmas que forman parte del ámbito en el que se sitúan las aportaciones de esta tesis. El capítulo comienza con el diagnóstico en STI, cuya definición ha sido brevemente esbozada en el capítulo 1, y que representa una pieza fundamental para conseguir STI eficientes.

3.1. Los Sistemas Tutores Inteligentes

3.1.1. Breve evolución histórica

Es posible hablar de sistemas educativos por computador desde principios de los años 60. Las primeras aplicaciones existentes eran meras interfaces entre profesor y alumno que permitían la planificación de cursos, la monitorización de ayudas suministradas al alumno, y la construcción de tests de maestría (Barr y Feigenbaum, 1982). Estos primeros sistemas se conocen con el nombre de *Sistemas de Enseñanza Asistida por Ordenador* (SEAO). Sus principales características son las siguientes (Urretavizcaya, 2001):

- Los cursos son muy extensos.

66CAPÍTULO 3. EL DIAGNÓSTICO EN LOS SISTEMAS TUTORES INTELIGENTES

- No existen una comunicación fluida entre el sistema y el alumno.
- El conocimiento del cómo y el porqué se ejecutan las tareas de enseñanza están fusionados.
- Son sistemas hechos a medida.
- El conocimiento del que disponen no evoluciona, es decir, no se modifica con el tiempo.

Desde los primeros sistemas educativos hasta nuestros días, se ha producido una evolución que se ha caracterizado por introducir en mayor o menor medida técnicas de IA como motor de decisión en la instrucción. Esta evolución permite clasificar los sistemas educativos de la siguiente forma (Urretavizcaya, 2001):

- Los *programas lineales*, que se caracterizaban por ser inmutables. Estos sistemas muestran el conocimiento de forma lineal, no se permite cambiar el orden de enseñanza establecido por el diseñador. Tienen su origen en la teoría conductista de Skinner (1985), que propugnaba que las personas funcionan por estímulos, y que a igual estímulo igual respuesta. Según esta teoría, no se debe permitir cometer errores a los alumnos. Además, en este tipo de sistemas, no se tenían en cuenta la aptitud del sujeto.
- Los *programas ramificados*, que tenían un número fijo de temas, al igual que los programas lineales, pero se diferenciaban en la forma de actuar según la respuesta del alumno. Utilizan la técnica de reconocimiento de patrones (*pattern-matching*), que permite tratar las respuestas no sólo como totalmente correctas o incorrectas, sino también como aceptables o parcialmente aceptables. Además se desarrollaron los denominados *lenguajes de autor* para permitir y facilitar la creación de contenidos educativos (Wenger, 1987).
- Por último, en los *sistemas generativos* o *sistemas adaptativos*, se sigue una nueva filosofía educativa, según la cual, los alumnos aprenden mejor si se enfrentan a problemas de dificultad adecuada, que atendiendo a explicaciones sistemáticas. Es decir, estos sistemas actúan adaptando la enseñanza a las necesidades de los estudiantes. Aunque en áreas como la aritmética han dado buenos resultados, en otras la dificultad para generar problemas es mayor.

3.1.2. ¿Qué es un Sistema Tutor Inteligente?

Los STI (en inglés, *Intelligent Tutoring Systems*) surgen como resultado de aplicar técnicas de IA a los SEAO. Además de en la IA, estos sistemas se apoyan en otras dos áreas de conocimiento: la Psicología Cognitiva y la Investigación Educativa. Nacen como un intento de complementar (o incluso de suplir) la cada vez más difícil tarea de proporcionar a cada alumno una instrucción personalizada. Una definición completa de un STI es la proporcionada por Wenger (1987):

Un STI es un SEAO que utiliza técnicas de IA, principalmente para representar el conocimiento, y dirigir una estrategia de enseñanza; y que es capaz de comportarse como un experto, tanto en el dominio de conocimiento que enseña (mostrando al alumno cómo aplicar dicho conocimiento), como en el dominio pedagógico, donde es capaz de diagnosticar la situación en la que se encuentra el estudiante, y de acuerdo a ello, ofrecer una acción o solución que le permita progresar en el aprendizaje.

En definitiva, un STI es un sistema que debe responder a tres preguntas fundamentales: *¿Qué es lo que se enseña?, ¿A quién se enseña (esto es, ¿Cuáles son sus características?)*, y por último, *¿Cómo se lleva a cabo esa enseñanza?* Además, y para diferenciarse de los SEAC, los STI deben aplicar técnicas de IA. La aplicación de este tipo de técnicas queda reflejada, principalmente, en dos aspectos que deben tenerse en cuenta durante el desarrollo de un STI: (a) Deben proporcionar una enseñanza individualizada, en función de las necesidades del estudiante en cada momento. (b) El orden y plan de interacción entre alumno y sistema nunca debe estar predefinido.

Las características principales de los STI son las siguientes:

- El conocimiento del dominio está acotado y estructurado.
- La información que se tiene del estudiante permite orientar la instrucción. Esta información se obtiene a partir de diagnósticos del estado actual.
- La secuencia de enseñanza no es fija, sino que se adapta a las necesidades de los estudiantes.
- Estos sistemas se enfocan más como herramientas complementarias a la enseñanza/aprendizaje que permiten mejorar su calidad, en vez de ser sustitutivas.

3.1.3. Arquitectura de un STI

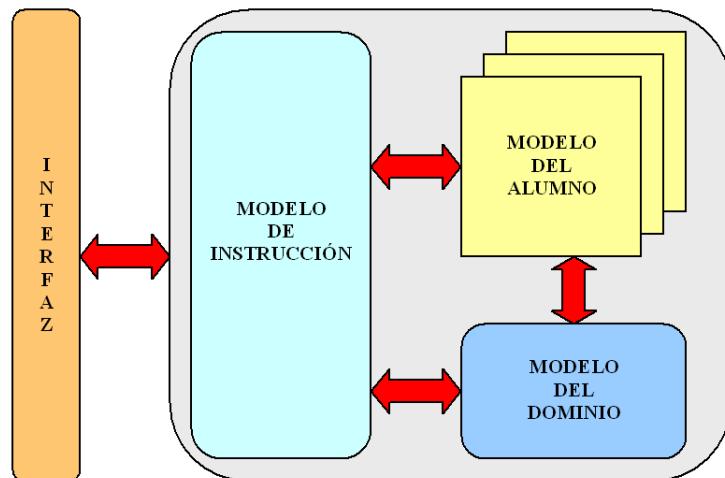


Figura 3.1: Arquitectura básica de un STI.

La figura 3.1 muestra la arquitectura básica de un STI, en la que se han representado sus componentes principales (Sleeman y Brown, 1982):

- *Modelo del dominio* (o Módulo experto): Corresponde a la respuesta sobre el *qué* se enseña. Contiene el conocimiento sobre la materia que debe ser aprendida. El primer paso en la implementación de un STI, es la representación explícita por parte del experto del conocimiento existente sobre el dominio. Un modelo del dominio será más

potente cuanto más conocimiento tenga (Anderson, 1988). Asimismo, es el encargado de la generación de problemas, y la evaluación de la corrección de las soluciones suministradas por el alumno. Su construcción requiere un esfuerzo notable de descubrimiento y codificación del conocimiento. Según Burns y Capps (1988) la lección más importante que la comunidad de investigadores en IA ha aprendido del desarrollo de sistemas expertos, es que cualquier modelo del dominio debe disponer de conocimiento abundante, específico y detallado, obtenido a partir de personas con años de experiencia en ese dominio. Los modelos del dominio pueden clasificarse a su vez de la siguiente forma (Anderson, 1988):

- *Modelo de caja negra* (en inglés, *black box model*): Es un medio de razonar sobre el dominio que no requiere una codificación explícita del conocimiento subyacente. Es decir, el modelo del dominio ha sido previamente codificado, aunque desde el punto de vista del sistema tutor no interesa cómo; lo único que interesa es conocer su comportamiento. Por ejemplo, si se deseara desarrollar un STI para enseñar operaciones matemáticas básicas, un buen modelo del dominio podría venir definido por una aplicación que llevara a cabo cálculos matemáticos.
- *Modelo basado en la metodología de los sistemas expertos* (en inglés, *expert system model methodology-based model*): Otra posibilidad es seguir los mismos pasos que en el desarrollo de un sistema experto. Esto supone extraer el conocimiento de un experto y decidir el modo en el que éste va a ser codificado y aplicado.
- *Modelo cognitivo* (en inglés, *cognitive model*): Supone hacer del modelo del dominio una abstracción del modo en el que los humanos hacen uso del conocimiento. Este tipo de modelos es el más efectivo desde el punto de vista pedagógico, aunque tiene la desventaja de que requiere un mayor esfuerzo de implementación. Su estructura estará condicionada al tipo de conocimiento que se quiera representar. Se pueden distinguir tres tipos:
 - a) *Conocimiento procedural*: es el conocimiento sobre el *cómo* realizar una tarea. En este caso, el modelo del dominio suele codificarse mediante un conjunto de reglas. Además, esta aproximación tiene la ventaja de que facilita la implementación de modelos de instrucción basados en el seguimiento de las acciones llevadas a cabo por el alumno.
 - b) *Conocimiento declarativo*: se limita a recopilar un conjunto de hechos y principios sobre el dominio y la relación entre éstos, es decir, a construir lo que se denomina *currículo*. Este tipo de modelos del dominio suelen ser representados mediante una *red semántica*, como un grafo acíclico en el que los nodos representan conceptos, que a su vez se enlazan entre sí mediante diversos tipos de relaciones: de agregación, de prerequisito, composición, etc.
 - c) *Conocimiento cualitativo*: es el conocimiento causal que permite a las personas razonar sobre comportamiento haciendo uso de modelos mentales.
- *Modelo del alumno*: El uso de modelos del alumno en STI surge como consecuencia del hecho de que estos sistemas deben trabajar con información incompleta, y por regla general con un alto grado de incertidumbre sobre los alumnos (Mayo y Mitrovic, 2001). Representa el a *quién* se enseña, es decir, lo que el alumno conoce y lo que no conoce del dominio. La mayoría de los STI infieren este modelo a partir de los conocimientos y carencias del alumno sobre el modelo del dominio, y a partir de esta información, adaptan el proceso de instrucción a sus necesidades. La estructura que almacena el estado de conocimiento del alumno es propiamente su modelo, mientras

que el proceso de razonamiento que actualiza este modelo se denomina *diagnóstico del alumno*. Según Burns y Capps (1988), el diagnóstico del alumno es una "aventura" de alto riesgo. La importancia de que éste sea certero es vital para el buen funcionamiento de un STI, ya que las estrategias tutoriales se deciden en función de la información que el sistema tiene sobre el estado en el que se encuentra el conocimiento del alumno.

- *Modelo de instrucción* (también llamado *modelo pedagógico* o *planificador de instrucción*): Corresponde al *cómo* se enseña. Constituye por tanto, las estrategias de enseñanza o estrategias tutoriales. Es decir, cómo el sistema debe mostrar el material educativo al alumno. Burns y Capps (1988) resaltan tres características tutoriales que debe tener un STI: (1) Control sobre la representación del conocimiento, para poder seleccionar y secuenciar las piezas que deben suministrarse al alumno. (2) Capacidad de responder a las preguntas de éste sobre objetivos de instrucción y contenido. (3) Estrategias para determinar cuándo necesita ayuda, y para seleccionar la ayuda más adecuada en cada momento.
- *Interfaz*: A través de ella se lleva a cabo la interacción hombre-máquina. Es necesario un esfuerzo adicional en el desarrollo de esta parte de la arquitectura, haciéndola intuitiva y transparente a los ojos del usuario alumno. Hay que tener presente que el estudiante no tiene porque ser una persona diestra en el uso de sistemas informáticos. Por este motivo, es importante que la interfaz sea fácil de manejar, ya que si no el alumno puede perder la concentración en el proceso de instrucción. Esto puede llegar a provocar que la sesión de instrucción no sea efectiva.

Como se puede apreciar, el desarrollo de un STI es un problema de gran magnitud. Por este motivo, es necesario hacer notar que en los STI existentes no todos los módulos anteriores están igualmente desarrollados (Barr y Feigenbaum, 1982). De hecho, muchos investigadores se limitan a desarrollar un único componente de un STI, que por sí solo, constituye un sistema independiente. Existen dos corrientes a la hora de desarrollar STI, en función de la estrategia de enseñanza seguida:

- *Enfoque instructivo*: Se centra en la transmisión de conocimiento entre el profesor y el alumno. El conocimiento debe estar bien estructurado. Debe disponerse de cierta variedad de técnicas para mantener la atención del alumno, y facilitar la transmisión del conocimiento. Se trata de una instrucción guiada, aunque el estudiante puede intervenir sugiriendo la realización de actividades.
- *Enfoque constructivo*: Se entiende el aprendizaje como un proceso activo de construcción de conocimiento (Piaget, 1952; Vygotskii, 1978; Minsky, 1986; Fostnot, 1996). Este planteamiento asegura además, que la mejor forma de aprender es dedicarse a construir de forma consciente algo, algún objeto. El alumno lleva el control de la actividad docente, construyendo su propia sesión de aprendizaje y fijando sus propios objetivos.

3.1.4. Los Sistemas Educativos Adaptativos para la Web

Los *sistemas hipermedia* son herramientas en las que el material docente está estructurado en páginas, estando éstas relacionadas entre sí mediante enlaces (hiperenlaces), que permite llegar de una a otra, y que permiten dotar a los cursos de una estructura jerárquica. Como resultado, el alumno puede navegar libremente por todo el contenido de un curso.

Este enfoque tiene como desventaja que la libertad del alumno puede convertirse en un arma contra sí mismo, ya que la movilidad por el hiperespacio puede contribuir a desorientarle.

El uso de Internet se ha extendido notablemente. Desde la década de los 90 han surgido diversos sistemas educativos que utilizan la web como medio de transmisión. Las ventajas de la utilización de la Web como medio son ampliamente conocidas: ubicuidad; disponibilidad inmediata; facilidad de instalación (el único requisito software para poder acceder es tener instalado un navegador web), etc.

Los *Sistemas Educativos Adaptativos para la Web* (SEAW) (Brusilovsky, 1998) (en inglés, *Web-based Adaptive Educational Systems*), representan un área de estudio dentro de los STI. Estos sistemas no son un nuevo tipo de sistemas educativos. Más bien, surgen de la conjunción de dos tipos de sistemas: los sistemas hipermedia adaptativos y los STI. En el estudio de los primeros, una rama de investigación más reciente (Brusilovsky et al., 1996a), se aplican modelos de usuario para adaptar al estudiante el contenido y los enlaces de las páginas. La educación es una de las aplicaciones de los sistemas hipermedia, pero no la única (sistemas de información personalizada, sistemas de recuperación de la información, etc.). En cuanto a los STI, bien es sabido que la adaptación es uno de los objetivos en su desarrollo, aunque ciertamente no el único. Generalmente, los SEAW se consideran STI implementados sobre la Web. Aunque existen pocos STI no implementados sobre la Web que utilicen técnicas de hipermedia adaptativa, la realidad es que casi todos los SEAW pueden clasificarse como STI además de como sistemas hipermedia adaptativos.

3.2. El modelado del alumno en los STI

El modelado del alumno es una de las piezas claves en el desarrollo de STI. Representa la creencia que tiene el sistema sobre cuál es el conocimiento del alumno (Holt et al., 1994). Para poder llevar a cabo una instrucción inteligente, adaptada a las necesidades del usuario, los STI mantienen un modelo del alumno. La calidad de la instrucción ofrecida por un STI, viene determinada por el alcance y la precisión de la información almacenada en el modelo del alumno, y por la habilidad del sistema para actualizarlo de forma dinámica. Como el modelo de alumno se utiliza como fuente de adaptación del sistema, en muchos casos, incluye información referente al comportamiento y conocimiento del usuario. Esta información tiene repercusión en su formación y conocimiento (Grigoriadou et al., 2002). Entre otras cosas, puede almacenar las características del alumno, su historial de navegación, etc. Además de esta información, el sistema suele mantener el nivel de conocimiento del individuo en cada uno de los conceptos del dominio.

La importancia del modelo del alumno en los STI es tal que una primera objeción a la utilización de estos sistemas es precisamente la posible no idoneidad de este modelo (Weber y Brusilovsky, 2001). Ciertamente es posible que la información recogida por el sistema sea insuficiente para modelar correctamente el conocimiento y las habilidades del alumno. Precisamente, algunos autores como Self (1994) han puesto de manifiesto la dificultad inherente al proceso de construcción de modelos del alumno (y en general, de usuario), debido al carácter intratable de esta tarea. A pesar de ello, estos autores reconocen que los modelos de usuario, aún no siendo completos desde el punto de vista cognitivo, ni de gran precisión, pueden ser de gran utilidad. La justificación de esta afirmación está en que los profesores, en general, utilizan modelos del alumno muy pobres, y aún así resultan ser más o menos efectivos.

En la construcción de modelos del alumno es necesario responder a cuatro preguntas (Stauffer, 1996): (1) *Quién*, es decir, el grado de especialización a la hora de definir a quién se está modelando. (2) *Qué*, esto es, lo que se va a modelar de ese alumno. Las opciones pueden ser: objetivos, planes, aptitudes, capacidades, conocimientos, creencias, etc. (3) *Cómo* se va a inferir y actualizar este modelo. Y por último, (4) *para qué*, es decir, cuál es el objetivo final de este modelo, cuál es su utilidad. Existen diversas alternativas: para suministrar refuerzos al alumno, para interpretar su conducta, para ayudarle, como medio delicitación de la información sobre su persona, etc.

Desde un punto de vista genérico, es posible generar modelos del alumno de dos formas: *Implícitamente*, monitorizando de forma pasiva y evaluando su comportamiento en la realización de tareas normales; o *explícitamente*, indicando al estudiante el camino que debe seguir en cada momento.

Verdejo (1994) aporta una clasificación bastante completa, considerando varios aspectos: (a) Según su *observabilidad*, pueden clasificarse en modelos externos, cuando el sistema ha sido desarrollado para un cierto perfil de alumnos, y por tanto, las decisiones de diseño están embebidas en el código del STI; por el contrario, los modelos internos son componentes separados, que permiten al sistema tratar explícitamente con las representaciones del alumno. (b) Según su *especialización*, un STI puede tener un único modelo del alumno, un conjunto de perfiles de modelos, o bien uno por cada estudiante. (c) Según el *número de modelos* por alumno. (d) Según su *variabilidad*, pueden definirse modelos estáticos si se definen una vez y no se modifican, o dinámicos, cuando el se modifica durante la instrucción. (e) Según su *ámbito temporal*, pueden construirse modelos a corto plazo, en los que éste sólo existe durante el periodo de instrucción, o a largo plazo, en los que se mantiene la información sobre el alumno después de la instrucción. (f) Según su *contenido*, éste puede ser conocimientos, intenciones, capacidades, preferencias y/o motivaciones. (g) Según su *extensión*, es decir, si abarcan sólo el conjunto de conceptos que el alumno conoce, o por el contrario, también aquellos que desconoce. (h) Según su *grado de conocimiento*, pueden clasificarse en modelos de conocimiento profundo, que son capaces de razonar sobre los modelos del dominio, y los de conocimiento superficial, que aunque utilizan técnicas para resolver un problema, no hacen uso explícito del modelo del dominio. (i) Según la *representación* utilizada, pueden ser inferenciales, que permiten simular el comportamiento del usuario, y no inferenciales, que se limitan a almacenar características a nivel descriptivo. (j) Según el *formalismo* utilizado en la implementación, esto es, lenguajes lógicos, representaciones basadas en marcos, etc. (k) Según la *técnica de construcción*, pueden ser implícitos, si es el propio sistema el que lo construye utilizando técnicas deductivas o inductivas, o bien explícitos, cuando son construidos bien sea por el diseñador del sistema y/o el propio alumno. (l) Según su uso, pueden clasificarse en descriptivos o predictivos.

Una extensión del trabajo de Verdejo (1994) fue la realizada por Holt et al. (1994), que ha dado lugar a la siguiente clasificación:

- *Modelos de recubrimiento o superposición* (en inglés, *overlay models*): En ellos se representa el conocimiento que el alumno tiene del dominio. Su comportamiento se compara, por tanto, con el de un experto. Las diferencias existentes se asumen como carencias en el conocimiento del alumno, que, de esta forma, se reduce a un subconjunto del que tiene el experto. La principal limitación de esta técnica de modelado está en su propia definición: el considerar que lo que el alumno sabe es únicamente un subconjunto del conocimiento del experto.
- *Modelos diferenciales*(en inglés, *differential models*): Dividen el conocimiento del alum-

no en dos categorías: el que se espera que el alumno tenga, y el que no se espera que tenga. Este tipo de modelos son una modificación de los de recubrimiento. La diferencia radica en que, en los diferenciales, las posibles carencias en el conocimiento no son todas igualmente deseables. Intentan representar de forma explícita las diferencias entre el conocimiento del alumno y el del experto. Además tienen la ventaja de que no son tan estrictos a la hora de modelar el conocimiento, aunque comparten las desventajas apuntadas en el modelo anterior.

- *Modelos de perturbación* (en inglés, *perturbation models*): En este caso, el conocimiento del alumno no es considerado únicamente un subconjunto de el del experto, sino que se contempla la posibilidad de que el estudiante posea ciertos conocimientos diferentes en cantidad y en calidad con respecto al del experto. Un aspecto diferenciador es que se combina lo que el alumno sabe correctamente, equivalente a un modelo de recubrimiento, con la representación de conocimientos erróneos que posee (denominados en inglés *misperceptions*), y con los procedimientos o conductas erróneas que realiza (en inglés, *bugs*). Esto permite tener datos más exactos de cuánto sabe el alumno. El conjunto de errores de concepto y de procedimiento están almacenados en una *biblioteca de errores*. El modelo del alumno se irá actualizando en función de la presencia o ausencia de este tipo de errores.

Existen diversas teorías para la construcción de la *biblioteca de errores* (Wenger, 1987): (a) *Teorías enumerativas*, donde todos los errores se enumeran basándose en un análisis empírico de los cometidos por el alumno. (b) *Teorías generativas*, donde los errores se generan en función de un conjunto de errores preestablecido. (c) *Teorías reconstructivas*, que dado un error observado, tratan de averiguar qué error ha dado lugar a que ése se produzca. El principal problema del uso de una biblioteca de errores es su construcción y posterior mantenimiento. Suele ser necesario un análisis exhaustivo de un gran conjunto de interacciones alumno-experto.

Aunque la inserción de información sobre los errores del alumno enriquece su modelo, este procedimiento no está exento de problemas. La construcción y mantenimiento del modelo se convierten en una ardua y costosa tarea, por lo que algunos autores han cuestionado su utilidad.

- *Modelos basados en intervalos de confianza* (en inglés, *bounded models*): En este tipo de modelos se postula que no es necesario conocer el estado exacto del conocimiento del alumno; basta con mantener un intervalo de confianza sobre sus límites inferior y superior. Han sido implementados utilizando técnicas de aprendizaje automático. A partir del comportamiento del alumno, el sistema infiere los límites inferior y superior de su conocimiento. Basándose en el modelo experto, se generan predicciones y problemas para verificar esas predicciones. Este tipo de modelos son más tratables ya que, en vez de intentar modelar el conocimiento del alumno tal como es, tratan con información más imprecisa.
- *Modelos basados en restricciones* (en inglés, *constraint-based models*) (Ohlsson, 1994): El alumno es representado como un conjunto de restricciones sobre la representación correcta del conocimiento. Este modelo es una extensión de los modelos de recubrimiento, permitiendo un razonamiento más sofisticado sobre los conceptos del dominio, y no limitándose a determinar si el alumno conoce o no esos conceptos. Cada vez que se viola una restricción sobre el dominio, se tendrá que llevar a cabo una actualización del modelo. Desde el punto de vista computacional, este tipo de modelos son simples, no estableciendo ninguna estrategia tutorial particular.

- *Modelos difusos* (en inglés, *fuzzy diagnostic models*): Este tipo de modelos se basa en conjuntos borrosos. En ellos se utilizan procedimientos estadísticos para propagar cambios de variables locales (habilidades para medir resultados) en variables más globales (habilidades para utilizar equipos). La presencia (o ausencia) de variables de conocimiento se representa mediante distribuciones de probabilidad con cinco niveles que van desde "ausencia total de conocimiento" hasta "conocimiento completamente desarrollado".

Por otro lado, Anderson et al. (1995) identifican dos tipos de técnicas de modelado del alumno, que se corresponderían con la clasificación de Verdejo (*op. cit.*) basada en la representación:

- *Evaluación o seguimiento del conocimiento* (en inglés, *assessment or knowledge tracing*): Consiste en determinar lo que el alumno sabe. Esto es, qué conceptos del modelo del dominio conoce, y cuáles han aprendido erróneamente. Este tipo de modelos son útiles para tomar decisiones pedagógicas, es decir, para decidir cuál es la siguiente acción que el estudiante debe llevar a cabo.
- *Identificación de objetivos o seguimiento del modelo* (en inglés, *plan recognition or model tracing*): Supone llevar a cabo un seguimiento o traza de la forma en la que el alumno soluciona un problema. Son útiles en STI cuyo objetivo es suministrar refuerzos, dar pistas o resolver dudas del alumno durante la resolución de un problema.

Hasta la aparición de las redes bayesianas, los STI hacían uso de razonamiento heurístico (VanLehn, 1988). Las *redes bayesianas* (Pearl, 1988), también llamadas *redes de creencia*, permiten realizar razonamientos probabilísticos en sistemas complejos de relaciones entre datos y resultados. Son grafos dirigidos en los que sus nodos representan variables aleatorias, y donde cada nodo tiene un número de estados.

Mayo y Mitrovic (2001) realizaron un estudio sobre el uso de redes bayesianas en el modelado del alumno en STI. En este trabajo, estos modelos se clasifican en tres tipos:

- En los *modelos basados en el experto*, es éste quien, a partir del modelo del dominio, define la estructura de la red bayesiana y establece los valores de las probabilidades condicionadas.
- En los *modelos basados en la eficiencia*, el experto especifica parcialmente la red y el conocimiento sobre el dominio es ajustado al modelo. La justificación de este tipo de modelos es maximizar la eficiencia de alguna forma (reducción del número de probabilidades a priori que es necesario especificar, reducción del tiempo de evaluación, etc.).
- Por último, en los *modelos basados en los datos*, la estructura y las probabilidades condicionales de la red son aprendidas a partir de un conjunto de datos obtenidos previamente.

Aunque los modelos del alumno fueron inicialmente concebidos para ser generados, actualizados y consultados única y exclusivamente por los restantes componentes de la arquitectura de un STI, la tendencia actual es el desarrollo de los denominados *modelos del alumno abiertos* (Dimitrova et al., 2001), en los que, durante el proceso de instrucción, el propio alumno puede consultar su modelo. Diversos autores han puesto de manifiesto las

ventajas que aporta el uso de modelos del alumno abiertos (Mitrovic y Martin, 2002; Mazza, 2003). Asimismo, mediante el uso de los denominados *modelos del alumno inspeccionables y/o modificables* (Bull y Pain, 1995), o también llamados modelos escrutables (Kay, 1995, 2000), se ofrece la posibilidad al alumno de no sólo consultar su modelo sino también, en ocasiones, de modificarlo. Weber y Brusilovsky (2001) justifican la utilización de este tipo de modelos para su STI ELM-ART. Según ellos, los sistemas instructores para la WWW pueden ser utilizados por estudiantes con un cierto conocimiento a priori del dominio, cuyo objetivo es refrescar o ampliar su conocimiento. Incluso algunos alumnos pueden tener experiencia en otros dominios relacionados y por lo tanto, tienen cierto conocimiento inicial por analogía. De esta forma, ellos mismos podrían inicializar su modelo, orientando el proceso de instrucción hacia aquellas áreas que no han sido exploradas, o en las que su grado de conocimiento es menor.

3.3. El diagnóstico del alumno en los STI

El objetivo principal de un sistema educativo es que el alumno aprenda nuevos conceptos y, como consecuencia, que su conocimiento y comprensión del dominio aprendido se vean incrementados. Por esta razón, el modelo del alumno debe actualizarse para reflejar los cambios en el estado de su conocimiento (Kavcic et al., 2002). Debido a que el canal de comunicación que se establece entre un estudiante y un STI es muy restrictivo, el STI únicamente es capaz de medir el conocimiento de forma directa, mediante la monitorización de la interacción con el alumno. El proceso de inferencia de las características internas del estudiante a partir de la observación de su comportamiento se denomina *diagnóstico del alumno* (VanLehn, 1988). Los aspectos fundamentales del diagnóstico del alumno hacen referencia a: (a) todas aquellas características observables del alumno que se almacenan en función de medidas específicas; (b) las características internas que deben ser inferidas en base a la información almacenada, y que son importantes para el aprendizaje; y (c) el método utilizado para extraer esa información a través de la monitorización y el seguimiento del sujeto.

La presencia de *incertidumbre* es un factor importante que frecuentemente lleva a errores en el diagnóstico del alumno. Esta incertidumbre aparece como resultado, en parte, de los errores y aproximaciones durante el proceso de análisis de los datos medidos, o bien es debida a la naturaleza abstracta de la percepción humana y a la pérdida de información resultante de su cuantificación (Grigoriadou et al., 2002).

Desde el punto de vista de la IA, la principal demanda exigida a un sistema de diagnóstico del alumno es el desarrollo de un método fiable y similar a la forma en la que lo haría un profesor. Este método debería ser capaz de analizar con efectividad, las medidas del comportamiento del estudiante. A partir de éstas, se hacen estimaciones sobre sus características internas, actualizando el modelo del alumno de acuerdo con esto. El problema está en que los sistemas de diagnóstico desarrollados se fundamentan en el uso de heurísticos. Como consecuencia, los resultados obtenidos son imprecisos y carentes de rigor. Por otro lado, los sistemas que hacen uso de métodos con un trasfondo teórico, proponen paradigmas poco viables desde el punto de vista práctico, cuya implementación suele tener requisitos difíciles de satisfacer. Otra desventaja adicional de los sistemas de diagnóstico es, en general, que se aplican a dominios muy concretos, siendo difícilmente extrapolables a otros diferentes.

Asimismo, otro de los problemas del diagnóstico del alumno es la *determinación inicial del estado de conocimiento*. Idealmente, sería necesario utilizar un método de inferencia que

permitiera, tras una breve interacción, obtener una primera medida de su conocimiento del dominio. El problema principal reside en que las técnicas bien fundamentadas que se utilizan suelen requerir bastante tiempo. Una posible solución a este problema de inicialización del modelo es el uso de *pretests*, que son tests que se llevan a cabo antes de que el estudiante comience su instrucción en un STI, con el objetivo de determinar su nivel de conocimiento inicial. Entre las ventajas del uso de pretests, Gouli et al. (2002) destacan las siguientes:

- Permiten inferir el conocimiento a priori del alumno en el dominio.
- Facilitan el diagnóstico de carencias en el conocimiento del alumno sobre otros dominios que representen prerrequisitos del dominio a estudiar.
- Permiten inicializar el modelo del alumno.
- Facilitan al estudiante una visión previa de la materia objeto de estudio.

3.4. Técnicas para el diagnóstico del alumno en STI

Esta sección estudia algunos de los STI o SEAW desarrollados por otros investigadores que de alguna forma utilizan tests. El objetivo principal es ver cómo realizan el modelado del alumno y sobre todo las herramientas de diagnóstico de las que disponen para actualizar esos modelos, así como determinar el papel que juega el uso de tests en estos sistemas. Adicionalmente, se analizarán otras herramientas y algoritmos para el diagnóstico y modelado del alumno (que también involucran de alguna forma tests) haciendo hincapié en las técnicas utilizadas para la actualización e inicialización de esos modelos.

Los sistemas estudiados se han dividido en cuatro familias, en función de cómo se lleva a cabo el diagnóstico del conocimiento del alumno: (1) basados en heurísticos, (2) basados en TAI, (3) basados en lógica difusa, y (4) basados en redes bayesianas. Al final de la sección se realizará un estudio comparativo de estos sistemas, destacando sus ventajas e inconvenientes.

3.4.1. Modelos de evaluación basados en heurísticos

En este apartado se recopilan todos aquellos sistemas que utilizan técnicas de evaluación no sujetas a ningún tipo de fundamento teórico, es decir, basados en heurísticos.

ELM-ART

ELM-ART (*Episodic Learner Model Adaptive Remote Tutor*) (Weber y Specht, 1997; Weber y Brusilovsky, 2001) es un sistema que funciona a través de Internet para enseñar conceptos básicos del lenguaje de programación LISP. Utiliza algunas de las ideas implementadas en otras aplicaciones anteriores:

- *Interbook* (Brusilovsky et al., 1996b), una herramienta para la generación de libros electrónicos.

- *ELM-PE* (Brusilovsky y Weber, 1996), un entorno inteligente de enseñanza que ofrece programación basada en ejemplos, análisis inteligente de soluciones a problemas y utilidades de depuración y prueba. Las características inteligentes se basan en el modelo ELM (Weber, 1996).
- *NetCoach* (Weber et al., 2001), es un sistema en alemán para crear cursos adaptativos a través de la web.

El *modelo del dominio* de ELM-ART está organizado en un red de conceptos, estructurada jerárquicamente en capítulos, secciones y subsecciones. Estas últimas, a su vez, se descomponen en páginas terminales o unidades. El sistema genera y actualiza un *modelo del alumno* abierto, inspecionable e incluso modificable por él mismo. Se trata de un modelo de superposición con los siguientes niveles o capas:

- *modelo visitado*, indica si el alumno ha visitado cierta página;
- *modelo aprendido*, actualizado a partir de evaluaciones realizadas a través de ítems o ejercicios resueltos por el alumno;
- *modelo inferido*, unidades que no han sido expuestas de forma directa al alumno pero que, por inferencias, se determina que el alumno tiene cierto conocimiento sobre ellas;
- y *modelo conocido*, si el alumno personalmente ha marcado que conoce cierta unidad.

A partir de la segunda versión de este sistema, el estudiante puede completar su formación mediante la realización de tests y ejercicios. Según sus autores, ELM-ART fue uno de los primeros sistemas inteligentes educativos que incluyó, como parte de su arquitectura, un componente para la realización de tests. Soporta cinco tipos diferentes de ítems: verdadero/falso; de opción múltiple, en los que el alumno debe seleccionar forzosamente una única respuesta; de respuesta múltiple, en los que el estudiante puede elegir una o más respuestas correctas; de respuesta libre; y de completar, donde hay que llenar un conjunto de huecos que aparecen en cierta sentencia. Entre otra información, junto con cada test se almacena su *dificultad*, definida como *"la cantidad de evidencia que es añadida al valor de confianza de los conceptos relacionados cuando se responde al ítem de forma correcta"*. También se almacena un *refuerzo* para los casos en los que el alumno responda de forma incorrecta, y el conjunto de conceptos relacionados con el ítem. Los ítems forman *grupos* que dan lugar a colecciones asociadas a una unidad específica, ofreciéndose la posibilidad de que un ítem pertenezca a varias con un peso asociado a cada uno de ellas. Si un ítem ha sido mostrado al alumno en un grupo, no volverá a presentarse en otros.

El alumno aprueba una unidad si el *factor de confianza* de ésta ha sido alcanzado o sobrepasado. Este umbral se define al especificar los parámetros del grupo. La evaluación del alumno j-ésimo, esto es, el cálculo de su factor de confianza c_j , se lleva a cabo mediante un heurístico, aplicado a los n ítems que ha respondido, de la siguiente forma: Por cada ítem i respondido de forma correcta ($u_{ji} = 1$), se multiplica su peso w_{ig} en el grupo g por su dificultad d_i , y el resultado se suma al factor de confianza. En caso de error ($u_{ji} = 0$), ese producto se multiplica a su vez por un factor de error e y se resta al factor de confianza.

$$c_j = \sum_{i=1}^n (-e)^{1-u_{ji}} w_{ig} d_i \quad (3.1)$$

Este sistema es, dentro de los STI, uno de los que hacen más uso de tests dentro de su funcionamiento habitual. Los tests que incorporan, intentan emular el funcionamiento de un TAI, es decir, en función de la respuesta del alumno, el siguiente ítem que se selecciona será más fácil o más difícil. El problema principal es que esta *dificultad* es un factor calculado sin ningún tipo de fundamento y basado únicamente en los criterios decididos por sus autores. El conocimiento del alumno se mide a través del denominado *factor de confianza*, cuyo valor se calcula aplicando de nuevo un heurístico (ecuación 3.1). Los términos de este heurístico son: (a) el peso del ítem en el grupo al que pertenece, que no es más que una forma de ponderar su importancia; (b) un factor de error, cuyo valor no queda muy claro cómo se calcula; y (c) la dificultad.

Además de la evaluación de una unidad, la realización de un test de grupo puede influir en la evaluación de otros conceptos. De igual forma que se actualiza el factor de confianza de la unidad, análogamente se modificarán los de todas las unidades con las que el ítem esté relacionado.

En ELM-ART, los ítems se presentan a los alumnos en tres situaciones:

- *Ejercicios*: Son presentados al final de una unidad, durante el aprendizaje de nuevos conceptos. Esto permite al propio alumno saber si ha asimilado un concepto, y a su vez proporciona información al sistema para actualizar el modelo del alumno. Este último tendrá que seguir resolviendo ítems mientras que el factor de confianza no alcance el nivel requerido para considerar el concepto evaluado como superado. La selección del ítem se realiza de la siguiente forma: Inicialmente se muestra al estudiante uno de dificultad media. En caso de error, el sistema elegirá aleatoriamente un ítem de entre aquéllos con dificultad menor. En caso de acertar, la selección se hará entre los de mayor dificultad.
- *Tests finales*: Son análogos a los anteriores, pero en este caso, se muestran al final de una lección, sección o subsección. En ellos, los ítems son presentados en grupos de entre 6 y 10.
- *Tests introductorios*: Permiten inicializar el modelo del alumno. Este último, al comienzo de una lección, sección o subsección, puede decidir empezar realizando un test que demuestre su conocimiento a priori sobre los conceptos que va a estudiar. En este caso, los ítems utilizados son todos aquéllos pertenecientes a elementos de menor nivel en el modelo del dominio. El tamaño máximo de un test de este tipo está restringido a 25 ítems, con un máximo de 5 por concepto. La evaluación en este caso es diferente: El sistema asume que para que el estudiante demuestre tener conocimiento suficiente de los conceptos relacionados, debe responder correctamente a todas las preguntas del test. Como resultado, todos los factores de confianza del modelo del alumno en esos conceptos se actualizan automáticamente a un valor superior al umbral.

Por otro lado, tanto en los ejercicios como en los tests finales, parece que no se finaliza hasta que el factor de confianza del estudiante no alcanza el nivel requerido. Esto puede traducirse en que se dé el caso de que alumnos con poco nivel se vean sometidos a sesiones excesivamente largas, que incluso podrían llegar a utilizar todos los ítems disponibles en el sistema.

Otra característica de la que carece este sistema es que, a pesar de definir pretests (tests introductorios) para inicializar el modelo del alumno, éstos sólo se aplican para inferir el conocimiento en un elemento del currículo del dominio (lecciones, secciones, ...)

no permitiendo una inicialización, como ocurre en el modelo presentado en esta tesis. Asimismo estos tests son muy restrictivos, ya que para que el sistema considere que el alumno sabe un concepto, debe responder correctamente a todos sus ítems.

DCG

DCG (*Dynamic Course Generation*) (Vassileva, 1997) es una herramienta para la creación de STI a través de la Web. Permite la generación de cursos individualizados en función de los objetivos de aprendizaje y del conocimiento previo del alumno. Separa la estructura del modelo del dominio del material pedagógico. Asimismo, adapta dinámicamente el curso de acuerdo con los logros del estudiante. Dado un concepto, objeto de estudio por parte del alumno, y su modelo de usuario inicializado a través de un pretest, el planificador de instrucción busca los subgrafos del modelo del dominio que conectan los objetivos con los nodos de éste. El modelo del alumno es de superposición sobre el del dominio. El estudiante, durante el proceso de instrucción, puede ser evaluado en cualquier momento mediante la realización de un test. Para calcular su nivel de conocimiento, tras realizar el test, se utiliza un heurístico. Si el alumno no tiene el nivel de conocimiento requerido en el concepto que acaba de estudiar, antes de avanzar y estudiar otros, se vuelve a mostrar el mismo pero esta vez utilizando otro material educativo. Si vuelve a no superar el nivel necesario se llevará a cabo una replanificación.

El sistema incluye una herramienta de autor para la construcción del modelo del dominio y para la inserción de los ítems. Cada concepto tiene asociado un fichero HTML con el material educativo y applets para incluir los siguientes tipos de ítems: opción múltiple, de respuesta libre y de ordenación de elementos. Por cada concepto, debe definirse el material educativo asociado y los ítems. Cuando se añade un ítem, debe especificarse su dificultad y un coeficiente que representa la contribución de una respuesta (correcta o incorrecta) a la puntuación global de los conceptos relacionados dentro del modelo del alumno. El nivel de conocimiento se define como el grado de conocimiento que tiene el estudiante sobre un concepto y está representado por una estimación probabilística. Para calcularlo se utiliza una fórmula (no facilitada por los autores) que tiene en cuenta el número y la dificultad de los ítems del tests correctamente resueltos, relacionados con el concepto.

DCG por sí solo no es capaz de decidir cómo presentar el material educativo. Para solventar esta deficiencia, se integra con la arquitectura GTE (Marcke, 1991), para llevar a cabo la presentación del material educativo. Además supone la inclusión en el modelo del alumno de sus preferencias (inteligencia, confianza, motivación, concentración, etc.) cuyos valores son asignados por el propio estudiante.

En este sistema, al igual que el ELM-ART, el modelo del alumno se inicializa administrando un pretest. Utiliza un procedimiento de evaluación basado en heurísticos, a partir de la dificultad del ítem y de un peso que determina su importancia para ese tema. La autora no describe cómo se calculan esos parámetros, pero es bastante probable que sea el propio profesor quien añada el material, el encargado de decidir estos valores. Ciertamente, los valores que proporciona un profesor pueden servir de orientación para el cálculo de la dificultad, pero nunca deben asumirse directamente sin aplicar ningún procedimiento adicional. Asimismo, tampoco se describe cómo se consigue que los tests sean de contenido balanceado en los pretest, ni el número de ítems que estos tienen, ni por último, cómo se seleccionan los ítems que se van a mostrar en el test.

ActiveMath

Es un sistema de enseñanza a través de la web que genera dinámicamente cursos de Matemáticas adaptados a los objetivos, preferencias, capacidades y necesidades de los alumnos (Melis et al., 2001). El conocimiento se estructura en conceptos, que pueden ser definiciones, axiomas, asertos, métodos de prueba, algoritmos, etc. A su vez, pueden estar relacionados con ítems, y pueden ser ejemplos, ejercicios, elaboraciones, motivaciones, o introducciones a nuevos contenidos. Los conceptos se enlazan entre sí mediante relaciones del tipo: dependencia matemática, prerequisitos pedagógicos, referencias; y los ítems se asocian a los conceptos a través de relaciones del tipo: ejemplo, ejercicio, motivación o prueba de un aserto.

Dentro del modelo del alumno se incluyen sus preferencias, esto es sus objetivos de aprendizaje y escenario que éste ha seleccionado: examen, preparación para un examen, resumen, resumen detallado, curso guiado o curso guiado detallado. Estos datos se extraen de un cuestionario que cada estudiante debe llenar la primera vez que accede al sistema. Además, éste debe realizar una estimación personal de su propio nivel de conocimiento en cada concepto del currículo del modelo del dominio, según tres niveles: bajo (color rojo), medio (color amarillo) o alto (color verde). Otra información requerida a priori es su destreza en la utilización de las herramientas externas integradas en ActiveMath que se utilizan en la presentación de los ejercicios.

Por cada concepto evaluado, el modelo del alumno almacena tres componentes: el *conocimiento* (relativo a lo que ha adquirido a través de la lectura), la *comprensión* (si ha seguido algún ejemplo) y la *aplicación* (si ha sido capaz de resolver correctamente algún ejercicio) en ese concepto. Estos componentes se han extraído de la taxonomía propuesta por Bloom (1956). Para actualizar estos valores, se han desarrollado dos herramientas de actualización del modelo: un *actualizador incremental*, que se limita a añadir una cantidad fija al elemento del concepto correspondiente; o un *actualizador bayesiano*, que utiliza una red bayesiana para llevar a cabo la actualización en función de dependencias condicionales.

Para llevar a cabo la adaptación, se utiliza toda la información almacenada en el modelo del alumno, junto con un conjunto de reglas pedagógicas. Estas reglas, a través de una máquina de inferencia basada en JESS (Friedman-Hill, 1997), permiten decidir: a) qué información debe ser presentada al estudiante, b) qué ejercicios y ejemplos deben ser mostrados, c) si debe o no utilizarse un sistema externo para este fin, y d) el orden en el que debe aparecer la información en la página.

Como ya se ha mencionado, en este sistema la inicialización del modelo del alumno la debe realizar el propio estudiante. Para ello se utilizan los tres niveles de conocimiento descritos. Esta opción tiene dos problemas: el primero de ellos es que el currículo es excesivamente grande, con lo que la inicialización se convierte en una tarea aburrida y tediosa. Por otro lado, parece poco apropiado que sea él mismo quien tenga que, forzosamente juzgar su nivel de conocimiento, sin darle ninguna otra alternativa para que sea el sistema el que a partir de una prueba de evaluación infiera estos valores.

En cuanto a los tests, se realizan a nivel de concepto, utilizan heurísticos para llevar a cabo la evaluación, y son siempre de longitud fija. Los autores no explican cómo a partir del resultado de un alumno en un test se lleva a cabo la actualización de su modelo. Sólo citan la existencia de un actualizador heurístico y otro bayesiano, sin decir cómo funcionan ni, para el caso del bayesiano, cómo se infieren las probabilidades a priori. A favor de este sistema hay que decir que ha sabido aprovechar la existencia de herramientas matemáticas ya consolidadas para incluirlas como medio de corrección de cierto tipo de ejercicios.

TANGOW

TANGOW (en inglés, *Task-based Adaptive Learner Guidance on the Web*) (Carro, 2001; Carro et al., 2001) es un sistema que genera cursos adaptativos a través de Internet. Esta aplicación genera de forma dinámica documentos que se presentarán a los alumnos durante un curso, y que se componen a partir de fragmentos de contenido proporcionados por el diseñador del curso.

En TANGOW, cada curso viene descrito mediante un conjunto de *tareas* y *reglas docentes*. Las *tareas* representan las unidades en las que puede dividirse el curso, y las *reglas docentes* determinan las relaciones entre éstas. La adaptación de los contenidos de un curso la lleva a cabo el denominado *Gestor de Tareas*. Su objetivo principal es decidir en cada momento la siguiente tarea o tareas que el alumno puede o debe realizar. Esto se hace en función de: (1) la estructura del curso, esto es las tareas y reglas definidas por el diseñador; (2) la estrategia de aprendizaje, seleccionada inicialmente por el estudiante (hay dos posibilidades: teoría antes de práctica o práctica antes de teoría); (3) sus datos personales, es decir, sus preferencias; y (4) las acciones que ha llevado a cabo con anterioridad durante la realización del curso. Las preferencias y datos sobre el perfil del estudiante se obtienen a través de un test que éste resuelve la primera vez que accede a un curso. Además de estos datos estáticos, el modelo del alumno almacena información dinámica obtenida a partir de la interacción con él, como por ejemplo el tiempo de dedicación a la realización de cada tarea, el número de páginas de teoría y ejemplos visitados, el de ejercicios resueltos, el de los que han sido solucionados de forma correcta, porcentaje de éxito, etc. Para el cálculo del éxito del estudiante, se utilizan los siguientes heurísticos, en función de si la tarea tiene un carácter práctico, teórico o mixto. Este valor se almacena en el modelo del alumno.

$$Exito_{practica} = \frac{ejerHechos}{totalEjer} \frac{ejerCorrectos}{ejerHechos} 100 \quad (3.2)$$

$$Exito_{teoria} = \frac{pagVisitadas}{totalPag} 100 \quad (3.3)$$

$$Exito_{compuesta} = [\frac{ejerHechos}{totalEjer} \frac{ejerCorrectos}{ejerHechos} 95] + [\frac{pagVisitadas}{totalPag} 5] \quad (3.4)$$

En esta herramienta, los factores de éxito para las tareas que involucran la realización de un test (esto es, práctica y compuesto) son una forma de representar el nivel de conocimiento del alumno. Si se simplifican esas fórmulas, se puede apreciar que corresponden al clásico criterio de evaluación basado en el porcentaje de ejercicios correctos. Asimismo, aunque no se expresa con claridad, parece que los tests son de longitud fija, y que el orden en que se muestran al alumno está predeterminado de antemano.

En cuanto al pretest realizado para inicializar el modelo del alumno, no se aporta ninguna información sobre su funcionalidad. En general, aunque el sistema parece combinar características interesantes desde el punto de vista de la adaptación de contenidos, sus autores conceden poca importancia al diagnóstico del conocimiento del alumno, olvidando la premisa de que un diagnóstico certero es vital para conseguir una adaptación adecuada.

HyperTutor

Hypertutor (Pérez et al., 1995) es un SEAW cuya arquitectura se divide principalmente en dos partes: el *componente tutor*, responsable del comportamiento adaptativo y el *componente hipermedia*, que dota al sistema de las características hipermedia. La adaptación que realiza se basa en el comportamiento del alumno y en la estructuración pedagógica del dominio. El modelo del dominio está formado por documentos hipermedia conectados entre sí mediante enlaces. A su vez, cada documento se compone de nodos, es decir, la información que es presentada al alumno, y de enlaces, que unen esos componentes y que representan los caminos que pueden seguirse para llegar a otros nodos.

El sistema maneja tres tipos de perfiles de alumno: novato, medio y experto. La instrucción se adapta en función de éste considerando resultados de interacciones anteriores. La evaluación se realiza con ejercicios (ítems de opción múltiple) siguiéndose un criterio porcentual. Éstos pueden solicitarse explícitamente o son propuestos por HyperTutor durante el proceso de aprendizaje; se caracterizan por nivel de dificultad y son seleccionados mediante el procedimiento siguiente: si se ha fallado en una pregunta (puntuación menor que 5) se propone uno similar, si sigue fracasando se propone uno más fácil y, en el resto de los casos, se propone uno más complicado. Para ello, los ejercicios se clasifican en tres categorías: de novatos, con valores de 0 a 3; de nivel medio, entre 4 a 7 y, por último, para expertos con rango de 7 a 10. Los profesores son los responsables de asignar la dificultad a cada ítem, y de asociarlos al concepto correspondiente. Adicionalmente, junto con la corrección de los ejercicios se muestra un refuerzo.

En este sistema, al igual que en ELM-ART, se intenta emular heurísticamente el funcionamiento de un test adaptativo. El problema está en que, también en este caso, la dificultad de cada ítem la determina el profesor según su propia experiencia. Asimismo, los autores no exponen claramente qué criterio aplican para determinar cuándo deben finalizar los test, aunque lo más probable es que se trate de tests de longitud fija. Asimismo, aunque la inicialización del modelo del alumno parece que se basa en el uso de tres perfiles, no se pone de manifiesto si HyperTutor es capaz de actualizar el perfil dinámicamente, a partir de sus resultados en los tests.

QUANTI

Este STI sobre Física Cuántica (Aimeur et al., 2001) se utiliza para enseñar cómo se procesa la información cuántica. Este dominio, según los autores, requiere del uso de modelos del alumno categorizados, ya que puede aplicarse a diversas disciplinas tales como la Informática, la Química, las Matemáticas o la Física. En el modelo del dominio, el conocimiento se representa utilizando redes semánticas, es decir, un grafo donde los nodos (entidades) son piezas de conocimiento y los vértices representan relaciones entre éstos. Esta red representa el nivel superior de la base de conocimiento. Cada concepto a su vez puede descomponerse en tres tipos de nodos: *componentes*, *características* y *ejemplos*. Un *componente* es una de las piezas de conocimiento que conforman el concepto; las *características* están asociadas a un concepto; y por último, los *ejemplos* sirven para ilustrar componentes.

El modelo del alumno se compone de tres submodelos:

- El *modelo cognitivo*, que se ocupa de representar el conocimiento del estudiante en el dominio. Se ha implementado mediante un modelo de superposición que deriva su estructura directamente del modelo del dominio, y en el que cada nodo es una pieza de conocimiento con un porcentaje que indica el nivel de conocimiento sobre el concepto.

- El *modelo afectivo* representa el estado emocional del alumno.
- El *modelo inferencial* representa las inferencias que se llevan a cabo a partir de la información de los otros dos modelos, los cuales se encarga además de actualizar.

Para inicializar el modelo cognitivo del alumno se utiliza una herramienta denominada CLARISSE (*Clusters and Rules Issued*) (Aimeur et al., 2002), que clasifica a los estudiantes dentro de un conjunto de *estereotipos* o categorías. El objetivo es, mediante la realización de un test con un número reducido de preguntas, determinar la categoría a la que pertenece el alumno y adaptar la instrucción en función ésta. Este resultado no es conocimiento a priori. Según los autores, la categorización puede considerarse como una forma de aprendizaje no supervisado, que puede definirse como la tarea de encontrar la estructura en los datos. Existen dos familias de métodos para este problema: unos contienen aproximaciones matemáticas y estadísticas (no ofrecen explicaciones sobre las inferencias obtenidas), y otros son aproximaciones simbólicas y conceptuales, como las utilizadas en CLARISSE, que ofrecen explicaciones sobre las creadas.

Para determinar el número de categorías en las que se pueden clasificar los alumnos, se utilizó una muestra de 31 individuos (profesores y estudiantes) de diversos países. Éstos fueron sometidos a un test de 30 ítems de opción múltiple de todos los temas enseñados en QUANTI. Las respuestas fueron puntuadas de tres formas: bien (10 puntos), mal (3 puntos) o muy mal (0 puntos). A partir de estos resultados, se aplicó un procedimiento de clusterización basado en el método del centroide, cuyo objetivo es identificar las distintas categorías de alumnos. Este procedimiento se describe con detalle en (Aimeur et al., 2002). El resultado fueron 7 diferentes categorías: (1) alumnos sin ningún tipo de conocimiento sobre la materia; (2) sujetos que saben de ciencias de la computación y algo de procesamiento de información cuántica; (3) estudiantes con amplios conocimientos de procesamiento información cuántica, pero que deben repasar; (4) alumnos con amplios conocimientos, aunque cometen fallos típicos; (5) individuos que saben mucho y no tienen errores; (6) con muchos conocimientos, pero que fallan cuestiones sobre computación universal; y (7) que saben mucho, pero deben revisar conceptos avanzados sobre procesamiento de información cuántica.

Para clasificar a nuevos alumnos se utiliza un test de ramificación fija con estructura en árbol, que consta únicamente de seis ítems. En función de cada respuesta, al sujeto se le mostrará un ítem u otro, de tal forma que con tan sólo tres ítems es posible determinar la categoría a la que pertenece.

Este trabajo utiliza un método interesante para la inicialización del modelo del alumno, basada en identificar la categoría a la que éste pertenece. La principal ventaja de esta técnica es que permite, con un número reducido de preguntas, determinar fácilmente la categoría a la que pertenece el alumno. Uno de los principales problemas que presenta es que las categorías fueron determinadas a partir de una muestra poblacional demasiado pequeña. De hecho, los propios autores reconocen esta limitación, y que los resultados obtenidos no presentan una fuerte correlación con el conocimiento real de cada individuo, tal y como sería de esperar. Asimismo, la forma de clasificar las respuestas a las preguntas es muy restrictiva (quizás por el hecho de que la muestra poblacional que han utilizado era muy pequeña): sólo se establecen tres niveles (bien, mal y muy mal) y dos de ellos son negativos.

Otra característica interesante del sistema es el uso de tests ramificados (véase información sobre este tipo de tests en la sección 2.9.1 del capítulo anterior). Trabajan con un árbol binario que, a partir de la respuesta del alumno (correcta o incorrecta), decide cuál es la siguiente pregunta que debe formularse, hasta que finalmente se determina la categoría a la que pertenece el alumno.

3.4.2. Modelos de evaluación basados en tests adaptativos

En el capítulo anterior se introdujeron los fundamentos de los TAI como mecanismo de evaluación bien fundamentado. En el campo de los tests adaptativos, la mayoría de las herramientas software existentes realizan un aproximación ecléctica de la estimación del conocimiento del alumno: no existe ningún estándar (ni siquiera un estándar de facto) (van der Linden y Pashley, 2001). En esta sección, se estudian varios modelos de evaluación que se basan en el uso de tests. En general, todos presentan algún tipo de adaptación, la cual no se fundamenta (al menos completamente) en el uso de heurísticos.

CBAT-2

Huang (1996a, 1996b) describe un algoritmo, CBAT-2 (en inglés, *Content-balanced Adaptive Testing*), para tests adaptativos en sistemas de enseñanza computerizados. Este algoritmo genera cuestiones de tests cuyo mecanismo de selección asegura una evaluación balanceada de los diferentes conceptos involucrados. Asimismo, utiliza un procedimiento simple para el aprendizaje de los parámetros de los ítems. Las características de CBAT-2 son las siguientes:

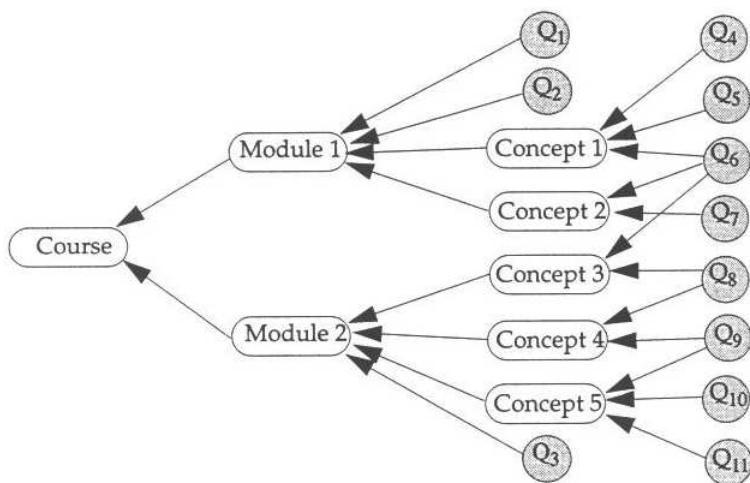


Figura 3.2: Estructura curricular utilizada por CBAT-2 (extraído de (Huang, 1996a)).

- *Contenido balanceado*: Asegura que los ítems seleccionados cubran de forma equilibrada todas las áreas de contenido evaluadas en el test, que se denominan *componentes*, y se estructuran jerárquicamente formando un gráfico acíclico y dando lugar a un currículo, tal y como se muestra en la figura 3.2. La raíz de la jerarquía es el *curso*. Cada componente puede tener cero o más componentes que representan sus subáreas de contenido. Los *módulos* son subáreas del curso. Los *conceptos* son a su vez subáreas de un módulo. Una cuestión pueden estar asociada a más de un componente del currículo a cualquier nivel.
- *Calibración del banco de ítems*: Tras un supuesto estudio empírico, se relajan, utilizando heurísticos, los requisitos necesarios para llevar a cabo una calibración de los ítems. La técnica heurística utilizada permite además realizar esta tarea en línea.

- *Selección inteligente de ítems:* Selecciona aquellos ítems que suministran más información para incrementar la eficiencia y la precisión del test.
- *Seguridad:* Escoge aquellos ítems que no forman un patrón.
- *Cuestiones de múltiples áreas:* Se permite que una cuestión esté asociada a más de un área.
- *Dos niveles de evaluación:* Ofrece información únicamente sobre el componente evaluado y los dependen de él de forma directa. Así, cuando un test valora un módulo, sólo se seleccionarán las cuestiones asociadas al concepto en cuestión y sus hijos.

Como parte de la inicialización, el algoritmo CBAT-2 genera un subcurrículo para cada test. El diseñador debe especificar el peso de cada componente dentro del test. La suma de los pesos de todos los componentes deben sumar diez. Si el profesor no especifica los valores de éstos, por defecto se asume que todas las áreas tienen el mismo. Además, debe indicar el número mínimo de preguntas que deben ser formuladas en el test por cada componente. Éstas tienen dos parámetros: el nivel de dificultad y el factor de adivinanza, cuyo significado es el mismo que en la TRI. El factor de discriminación no se utiliza porque, según Huang, se trata de un valor complicado de medir y su significado es difícil de entender por los profesores. En cuanto al factor de adivinanza, se estima a partir del número de respuestas correctas sobre el total de respuestas, y su valor se mantiene constante. El valor de la dificultad oscila entre 0 y 1. Inicialmente el profesor se encarga de asignarle valor según su experiencia. Posteriormente, y conforme se vaya almacenando información de sesiones de alumnos, se modifica utilizando la siguiente fórmula:

$$dif_i = \frac{20init_i + \Phi_i}{20 + R_i + W_i} \quad (3.5)$$

donde $init_i$ es la dificultad inicial del ítem; la constante 20 un factor de normalización; R_i el número de veces que la pregunta ha sido respondida correctamente; y W_i el número de veces que la pregunta ha sido respondida incorrectamente. Φ_i es un acumulador de la dificultad para el ítem que se define de la siguiente forma:

$$\Phi_i = \sum_{j=1}^n k_j f(\Theta'_j) \quad (3.6)$$

siendo $n = R_i + W_i$ y Θ'_j el nivel de conocimiento provisional del alumno. Cuando la respuesta es correcta k_j toma el valor cero. Si por el contrario la respuesta es incorrecta, k_j toma el valor 2. La función f convierte un número en la escala $[-4, 4]$ en un nivel de dificultad entre $[0, 1]$.

El proceso de selección de ítems se compone de dos fases. En la primera se determina el área de contenido sobre la que se seleccionará el ítem. Para ello, se eligen un conjunto de componentes candidatos, entendiéndose éstos como aquéllos cuyo nivel de conocimiento no ha sido estimado todavía. La elección se lleva a cabo probabilísticamente. Para ello se calcula la probabilidad de selección del componente de la siguiente forma:

$$P_i = \frac{W_i}{\sum W_j | C_j \text{ es un componente}} \quad (3.7)$$

En la segunda fase, se selecciona de entre los ítems del componente que no han sido mostrados al alumno, aquél que suministre más información a la estimación de su conocimiento. La función de información se calcula utilizando, como CCI, la función logística de tres parámetros en la que el factor de discriminación es siempre constante e igual a 1,2. La elección de este valor se ha hecho en virtud de un estudio previo realizado por Kingsbury y Weiss (1979), en el cual se calibró un banco de ítems y la media de los factores de discriminación era 1,2. Para calcular el valor de la dificultad, cuyo valor debe oscilar entre 0 y 1, se aplica una función g inversa a la f aplicada en la ecuación 3.6.

En la fase de estimación del conocimiento del alumno, el algoritmo calcula el conocimiento sobre el componente padre del subcurrículo del test y de todos sus descendientes. En caso de que éste tenga asociadas preguntas, se crea además una componente ficticio al que se le asigna un peso. Cada vez que el sujeto responde a una pregunta, se actualiza su estimación de conocimiento global del test, y del componente que corresponda. Se utiliza el criterio bayesiano para hacer la estimación.

Si el nivel de confianza de la estimación global es mayor que un cierto umbral, y además se ha mostrado al examinando un número mayor de preguntas que el indicado por el profesor para todos los componentes del test, éste finaliza.

Este algoritmo ha sido utilizado para desarrollar varios sistemas para la administración de tests adaptativos (Gonçalves, 2004; Gonçalves et al., 2004). Se trata de un algoritmo sencillo de implementar en el que parecen resolverse dos de las principales problemas que presentan los TAI basados en la TRI. Éstos son la inexistencia de un mecanismo para asegurar tests de contenido balanceado, cuando en un mismo test concurren múltiples conceptos; y el problema del alto coste de la calibración inicial de los ítems.

La desventaja principal de CBAT-2 radica precisamente en cómo se han solucionado los problemas de los TAI. Los tests de contenido balanceado se consiguen gracias a la inclusión de pesos, determinados manualmente por los profesores. Esta solución heurística hace que los tests pierdan su rigor. Asimismo, este algoritmo sólo permite llevar a cabo evaluaciones en dos niveles. Esto limita su uso en tests que evalúen conceptos de jerarquías de más de dos niveles. Como se verá en capítulos posteriores, en esta tesis se propone un mecanismo que asegura tests de contenido balanceado sin necesidad de incluir ningún tipo de heurístico, y que permite evaluar tantos conceptos como se deseé, sin estar condicionado por el número de niveles de la jerarquía curricular.

En el capítulo anterior se mostraron las técnicas de calibración más usuales. Éstas requieren un número relativamente alto de sesiones de tests realizadas (de forma no adaptativa) con los ítems que se desea calibrar. En general, esta información inicial es costosa de conseguir. CBAT-2 incluye un método en el que se relajan estos requisitos iniciales, que además permite la calibración en línea. Esta solución se basa en heurísticos y, por tanto, en fórmulas carentes de fundamento teórico. Como resultado, lo más probable es que los ítems que se obtengan estén mal calibrados, lo que a su vez supone diagnósticos incorrectos. Uno de los objetivos de esta tesis es paliar, en la medida de lo posible, el problema de la calibración, proponiendo un método no heurístico, y además con menos requisitos iniciales que los algoritmos usados tradicionalmente para ello.

HEZINET

Es un SEAW para la enseñanza de la lengua vasca completamente desarrollado e implantado en diversos centros de educación (Gutiérrez et al., 2002; López Cuadrado et al., 2002, 2002). Este sistema integra una aplicación para la administración de TAI. Tienen un

banco con unos 600 ítems, y utilizan como modelo de respuesta la TRI dicotómica 3PL. Aunque en la actualidad, según sus autores, los ítems no están calibrados adecuadamente, el objetivo es llevar a cabo una calibración de los ítems administrándolos previamente de forma convencional, para luego poder integrarlos y utilizarlos como tests adaptativos.

Para la actualización del modelo del alumno, el sistema Hezinet utiliza cuatro tipos de tests: *tests de sesión*, sobre los contenidos que el alumno ha estudiado recientemente; *tests de capa*, que incluyen ítems que cubren el conocimiento de la parte que se está estudiando en ese momento; *tests de curso*, sobre los contenidos clave del curso que el examinando está realizando actualmente; y *tests de admisión*, que cubren todos los contenidos de todos los cursos del programa. Para inicializar el modelo del alumno y como nivel estimado inicial, son varias las propuestas que hacen (un valor aleatorio, una pequeña prueba previa para calcularlo, un valor indicado por el propio usuario, etc.), aunque los autores no exponen con claridad cuál es la que realmente piensan utilizar. Otros aspectos que no se ponen de manifiesto son: (a) cómo se asegura una selección de ítems balanceada en contenido; (b) si a través de un único test se permite la evaluación de múltiples conceptos; y (c) qué procedimiento de calibración de ítems va a utilizarse.

Modelo basado en la Teoría de la Decisión

Rudner (2002) hace una propuesta innovadora en el uso de la *Teoría de la Decisión* como modelo de respuesta en tests adaptativos. El modelo propuesto consta de los siguientes componentes: (a) Un conjunto de K niveles de conocimiento m_1, m_2, \dots, m_k , en los que se clasifica a los alumnos; (b) un conjunto de ítems calibrados; (c) los patrones de respuesta de una muestra de alumnos $\mathbf{z} = z_1, z_2, \dots, z_n$; y (d) el espacio de decisiones, esto es, las acciones que deben ser tomadas por los usuarios.

A priori, un test basado en este modelo, dispondría de una distribución de alumnos sobre los niveles de conocimiento $p(m_k)$, que podría calcularse a partir de un test administrado anteriormente u otra información que, en caso de no estar disponible, se asume que es uniforme. Otro elemento necesario es la probabilidad de que los alumnos respondan correctamente al ítem i dado su nivel de conocimiento $p(z_i|m_k)$, es decir, lo que en la TRI correspondería a las CCI.

Para el cálculo de la estimación del conocimiento del alumno se aplica el teorema de Bayes de la siguiente forma:

$$P(m_k|\mathbf{z}) = cP(\mathbf{z}|m_k)P(m_k) \quad (3.8)$$

donde c es la constante de normalización que asegura que la suma de probabilidades de la distribución sea 1.

$$c = \frac{1}{\sum_{k=1}^K P(\mathbf{z}|m_k)P(m_k)} \quad (3.9)$$

Se asume asimismo la independencia local de los ítems, con lo que $P(\mathbf{z}|m_k)$ se calcula como el producto de los $P(z_i|m_k)$. El nivel de conocimiento estimado será el de mayor probabilidad de la distribución. También se define la probabilidad de responder correctamente a un ítem i de la siguiente forma:

$$P(z_i = 1) = \sum_{j=1}^K P(z_i = 1|m_j)P(m_j) \quad (3.10)$$

Como mecanismo de selección de ítems se proponen tres criterios diferentes. Todos ellos se basan en el cálculo del *coste medio* o *coste esperado* B , que viene dado por:

$$B = \sum_{i=1}^K \sum_{j=1}^K c_{ij} P(m_i|\mathbf{z}) P(m_j|\mathbf{z}) \quad (3.11)$$

donde c_{ij} es definido de forma genérica como el coste de decidir d_i cuando m_j es verdadero, y toma por defecto el valor 1.

Los criterios son los siguientes:

- *Mínimo coste esperado.* Se basa en aplicar el siguiente procedimiento a cada ítem:
(1) Asumiendo que el alumno va a responder correctamente al ítem, se calculan la probabilidad a posteriori según la ecuación 3.8, y el coste correspondiente según la ecuación 3.11. (2) Se sigue el mismo procedimiento pero esta vez suponiendo que el alumno responde incorrectamente. (3) Se multiplica el coste obtenido en el paso 1 por la probabilidad de una respuesta correcta al ítem. (4) Se multiplica el coste obtenido en el paso 2 por la probabilidad de una respuesta incorrecta al ítem. (5) Se suman los valores obtenidos en los pasos 3 y 4. El ítem elegido es aquél con menor coste esperado.
- *Ganancia de información.* El objetivo en la selección de ítems es elegir el ítem que suministre mayor información sobre el nivel de conocimiento del alumno. La medida que se utiliza comúnmente en la Teoría de la Información es la *entropía*.

$$H(S) = \sum_{k=1}^K -p_k \log_2 p_k \quad (3.12)$$

La *entropía* es una medida de la uniformidad de una distribución. El objetivo es, por tanto, tener una distribución de $P(m_k)$ con forma apuntada, y seleccionar el ítem que supone la mayor disminución de la entropía, es decir, $H(S_0) - H(S_i)$ donde $H(S_0)$ es la entropía ponderada actual y $H(S_i)$ la de después de aplicar el ítem i . Su cálculo se realiza de la siguiente forma:

$$H(S_i) = p(z_i = 1)H(S_i|z_i = 1) + p(z_i = 0)H(S_i|z_i = 0) \quad (3.13)$$

El procedimiento para elegir el siguiente ítem en un test se basa en los siguientes pasos, que se aplicarán a todos los que no han sido administrados: (1) Se calculan las estimaciones a posteriori suponiendo respuesta correcta e incorrecta, y aplicando la ecuación 3.8. (2) Se calculan las entropías condicionales para ambas respuestas según la ecuación 3.12. (3) Se calcula la entropía ponderada según la ecuación 3.13. El ítem seleccionado es aquél que suponga una reducción mayor de la entropía.

- *Discriminación máxima.* Este criterio se basa en elegir aquel ítem que mejor discrimina entre los dos niveles de conocimiento más probables para la estimación actual. Un indicador de discriminación es el siguiente:

$$M_i = \left| \log \frac{p(z_i = 1|m_k)}{p(z_i = 1|m_{k+1})} \right| \quad (3.14)$$

donde m_k y m_{k+1} son los niveles de conocimiento más probables.

Aunque este modelo está restringido únicamente a ítems dicotómicos, las ventajas frente al uso de la TRI en tests adaptativos son destacadas, según Rudner, en aquéllos con un número pequeño de ítems. Estas ventajas se traducen en una mejora en la precisión de la estimación, uso de bancos de tamaño reducido, facilidad de implementación, etc. Para sustentar que este modelo se comporta mejor que los basados en la TRI, Rudner llevó a cabo un experimento con alumnos simulados en el que comparó la precisión de las estimaciones realizadas según ambos formalismos. El resultado de este estudio fue que, aunque para dos niveles de conocimiento con la TRI se obtenían resultados sensiblemente mejores, para cuatro niveles los resultados obtenidos por su propuesta eran mejores. Hay que resaltar que Rudner compara los resultados aplicando sus tres criterios de selección, con únicamente el criterio de selección basado en la máxima información de la TRI. Esto supone una clara desventaja de un modelo frente al otro.

Otra de las desventajas señalada por el propio autor, es que su modelo no contempla la realización de tests de múltiples conceptos de forma simultánea. Consecuentemente, no se dispone de ningún mecanismo para asegurar el balanceo de contenido en este tipo de tests.

En capítulos posteriores, se mostrará cómo, en esta tesis, se han aprovechado algunas de las ideas de Rudner, para elaborar mecanismos de selección de ítems basados en la TRI; y cómo estos nuevos criterios permiten el diagnóstico del alumno a través de un modelado polítómico de la respuesta a los ítems.

Evaluación con tests adaptativos en INSPIRE

INSPIRE (en inglés, *Intelligent System for Personalized Instruction in a Remote Environment*) (Gouli et al., 2001; Papanikolaou et al., 2003) es un SEAW que restringe el dominio de aprendizaje al comienzo de la instrucción, para ir extendiéndolo progresivamente según ésta va avanzando. Se basa en objetivos de aprendizaje escogidos por el alumno, conforme a los cuales se generan las lecciones. Otros parámetros del estudiante que influyen en la selección de las lecciones son su nivel de conocimiento y el estilo de aprendizaje que ha elegido. El modelo del dominio es jerárquico y está estructurado en tres niveles: objetivos de aprendizaje, conceptos y módulos de conocimiento. El mecanismo de adaptación se basa en tres principios: (a) la evaluación debe reflejar qué es lo más importante que el alumno debe aprender; (b) ésta debe mejorar el aprendizaje y ofrecer instrucción práctica; y (c) cada alumno tiene la oportunidad de aprender mientras es evaluado.

En este sistema, los tests adaptativos se utilizan para evaluar el progreso en el aprendizaje, y además proporcionan refuerzos al alumno en función sus respuestas. En el proceso de elicitation, el tutor especifica un conjunto de ítems, que a su vez asocia a los correspondientes módulos de conocimiento. Es posible tener un ítem para más de un módulo. Cada ítem tiene asociado un conjunto de parámetros: su nivel de dificultad, el número de veces que ha sido contestado correcta o incorrectamente, etc. Asimismo, el tutor define especificaciones

de tests, donde debe configurar entre otros parámetros: el número mínimo de preguntas que deben ser mostradas, el nivel de precisión requerido para poder finalizar el test, etc. El sistema ofrece al alumno dos posibilidades de evaluación:

- *Tests de autoevaluación*, sobre los módulos de conocimiento que ha estudiado, para que de esta forma, sea el propio alumno quien mida sus progresos. Se construyen en función de cuánto sabe y del historial de navegación a través de los módulos de conocimiento.
- *Tests de recapitulación*, sobre algún concepto recientemente estudiado, o bien sobre el objetivo de aprendizaje del alumno. En estos tests, no se tiene en cuenta el historial de navegación del mismo.

El modelo de respuesta utilizado se basa en la TRI. Para modelar la CCI se utiliza la función 3PL en la que el factor de discriminación permanece constante e igual a 2. El nivel de conocimiento oscila entre -3 (novato) y 3 (experto). Inicialmente, el profesor será el encargado de estimar el valor de la dificultad de cada ítem. Posteriormente, y conforme se vayan teniendo resultados de sesiones realizadas por alumnos, los ítems se irán calibrando adecuadamente siguiendo el heurístico propuesto en CBAT-2. Igualmente, el factor de adivinanza se fija (y no se actualiza) como el número de respuestas correctas dividido entre el total de posibles respuestas.

Tanto para la selección del siguiente ítem, como para la estimación del nivel de conocimiento, se ha elegido el criterio de máxima verosimilitud. A priori, el nivel de conocimiento del alumno es un nivel medio. En cuanto a los criterios de finalización del test, existen dos posibilidades: un número máximo de ítems mostrados al alumno, o si la precisión de la estimación excede un cierto umbral.

PASS

Desarrollado por el mismo grupo de investigación que ha desarrollado INSPIRE, PASS (en inglés, *Personalized Assessment Module*) (Gouli et al., 2002) fue concebido como un módulo de diagnóstico, útil para ser integrado en SEAW con modelos del dominio estructurados, y modelos del alumno que almacenen información sobre cuánto sabe éste, extraída a raíz de interacciones a través de la navegación. Como mecanismo de evaluación se utilizan tests y cuestiones, ambas adaptativas. Se definen tres tipos de evaluación para la ayuda a la toma de decisiones: pretests, tests de autoevaluación y evaluación recapitulativa. El material educativo de cada concepto está organizado en tres niveles de actuación: *nivel de recuerdo* (capacidad de recordar), *nivel de uso* (capacidad de aplicar los conceptos) y *nivel de búsqueda* (capacidad de proponer y resolver problemas), basándose en la taxonomía de Bloom (1956).

En los pretests, se incluyen cuestiones de tres categorías distintas: (a) Cuestiones para ofrecer al alumno una primera impresión sobre los contenidos que va a estudiar. Se utilizan para inferir su conocimiento a priori. (b) Cuestiones de respuesta corta sobre su experiencia y su conocimiento a priori sobre dominios relacionados. (c) Cuestiones sobre prerequisitos del dominio que se está estudiando. Inicialmente, el sistema selecciona dos categorías sobre las que realizar preguntas que pertenecerán a una determinada clase. Cuando la aplicación sea capaz de diagnosticar el conocimiento del estudiante en esa categoría, le propondrá preguntas de otra clase, y así sucesivamente. Para evaluar aplica un modelo cuantitativo, donde

90CAPÍTULO 3. EL DIAGNÓSTICO EN LOS SISTEMAS TUTORES INTELIGENTES

el alumno es clasificado en cuatro niveles. Una vez terminado el pretest, se estima el conocimiento del examinando en cada concepto, teniendo en cuenta: (a) El porcentaje de respuestas correctas en las preguntas sobre aquellos conceptos que va a estudiar. (b) El nivel de conocimiento en los conceptos que son prerequisitos. (c) Los pesos de los conceptos prerequeridos, y su importancia sobre aquéllos a los que preceden.

La construcción de tests para autoevaluación o para recapitulación se hace de forma dinámica, en función del conocimiento del alumno, teniendo en cuenta el número máximo de preguntas por nivel de actuación. Para el caso de autoevaluación, se tiene en cuenta también la navegación que ha realizado en el curso y el tiempo empleado en estudiar cada material.

El procedimiento de evaluación es el siguiente: se busca el nivel de conocimiento del alumno. Si no ha sido estimado, se toma un valor medio. En función del tipo de test, se selecciona su número máximo de preguntas. Para autoevaluación, se emplea el peso del material educativo y las páginas visitadas. Para los tests de recapitulación, únicamente el peso del material educativo. Las cuestiones candidatas se calculan en función del tipo de evaluación, el número máximo y mínimo por nivel de actuación, y las páginas visitadas y sus pesos. Para cada pregunta, se calcula su CCI y su función de información. Se selecciona aquélla que sea más informativa. Las curvas se calculan utilizando dos parámetros: el factor de adivinanza y la dificultad. El primero de ellos se calcula dividiendo uno entre el número de opciones, y el segundo se calibra de forma incremental utilizando el procedimiento definido en CBAT-2. Cada ítem almacena un refuerzo por cada respuesta, los parámetros de la CCI, el número de alternativas de respuesta y el nivel de actuación evaluado. Por cada test se guarda el número máximo de preguntas por cada nivel de actuación, el umbral de confianza para la evaluación, el criterio de finalización, el peso de cada material educativo y de cada prerequisito.

Para probar la utilidad de este sistema, los autores hicieron un pequeño estudio con veinte estudiantes, los cuales trabajaron durante dos horas con el sistema INSPIRE, utilizando PASS como módulo de evaluación. Los resultados obtenidos fueron que el sistema estima de la misma forma que lo haría un experto, en este caso, el profesor de la asignatura.

Aunque, según sus autores, el objetivo de PASS es ser una herramienta de diagnóstico independiente, no se explica con claridad qué requisitos debe cumplir el SEAW donde se integre PASS, ni cómo se lleva a cabo la comunicación entre ambos sistemas. Como se verá en capítulos posteriores, una de las aportaciones de esta tesis es un sistema para el diagnóstico también integrable en otros sistemas (con ciertas restricciones). Por este motivo, se ha definido un breve protocolo que determina los términos en los que se llevan a cabo las interacciones entre sistemas externos.

Asimismo, los autores de este sistema han determinado que la evaluación del alumno se realiza de forma discreta únicamente en cuatro niveles de conocimiento, sin justificar el porqué de esta elección. Aunque la implementación de sus tests en general se basa en la utilización del algoritmo CBAT-2, sus pretests no parecen utilizar el balanceo de conceptos, sino que se construyen mediante el encadenamiento de otros más pequeños sobre las distintas clases. Por otro lado, los autores dan algunos datos sobre cómo se estima el conocimiento del examinando, sin explicar de forma precisa cómo lo hacen.

SKATE

Esta propuesta (Chua Abdullah, 2003) para la construcción y actualización de modelos de usuario se basa en el uso de tests adaptativos en los que en vez de utilizar la TRI

como modelo subyacente, se utiliza la *Teoría del Espacio de Conocimiento (Knowledge Space Theory)* (Falmagne et al., 1990). En SKATE, el modelo del dominio está basado en restricciones, cuyo lenguaje de representación es PROLOG. A diferencia de la mayoría de sistemas que utilizan restricciones, en SKATE se representan las características de los problemas, en vez de representar reglas que permitan detectar errores. El sistema ha sido aplicado al dominio de la suma de quebrados, donde los problemas están organizados en nueve categorías. SKATE también permite la generación dinámica de problemas. Como mecanismo de diagnóstico, aplica tests en los que el orden en el que los ítems se muestran a los examinandos lo especifica manualmente el profesor por medio de un árbol de decisión, en función del número de habilidades que el alumno debe poseer para resolver cada problema. A la vista de esto, parece más adecuado llamarlo test con ramificación en vez de adaptativo.

3.4.3. Modelos de evaluación basados en lógica difusa

La *lógica difusa* (Zadeh, 1965) intenta modelar el razonamiento humano que de por sí es impreciso, incierto y ambiguo. Puede ser considerada como un enfoque para cuantificar grados de conocimiento, pero en un sentido más cercano al lenguaje natural (Millán, 2000). Para ello, utiliza tres tipos de elementos: variables, valores y conjuntos. Las *variables difusas* permiten definir conceptos cuyos valores no pueden ser completamente cuantificados. Los *valores difusos* representan al rango que puede tomar una variable. Por ejemplo, la variable difusa *temperatura* podría tomar los valores difusos *alta*, *media* o *baja*. Un *conjunto difuso* no es más que una función de pertenencia cuyo dominio son los valores difusos y el rango valores reales entre 0 y 1. Asimismo, es posible operar con los conjuntos difusos, mediante operadores como la negación, intersección, unión, etc... (Millán, 2000, cap. 2).

Tests adaptativos basados en conjuntos difusos

Neira et al. (2002) hacen una propuesta de un modelo basado en TAI que utiliza conjuntos difusos, en vez de la TRI. Se basa en la fusión de tres modelos conceptuales genéricos: el bayesiano de Owen (1975) para la selección de ítems y estimación del conocimiento del alumno; el de Birnbaum (1968); y el algoritmo CBAT-2 de Huang (1996a, 1996b) de construcción de TAI con contenido balanceado aplicados a modelos del dominio jerárquicos.

Se define un modelo de tests de contenido equilibrado en el que los ítems están asociados a los conceptos que evalúan, los cuales a su vez están estructurados de forma jerárquica. El modelo del dominio se estructura dividiendo las materias en *áreas* que están divididos en *unidades*, que a su vez contienen *conceptos*, los cuales se evalúan a través de *ítems*. Los alumnos son evaluados en los siguientes cinco niveles de conocimiento: Muy deficiente (M), Insuficiente (I), Suficiente (S), Notable (N) y muy Bien (B).

El proceso de selección del siguiente ítem del test se divide en dos fases. En un primera, se selecciona el área de contenido sobre el que se evaluará al alumno de entre un conjunto de candidatas. Un área es *candidata* si aún no se ha decidido de modo fiable la calificación del estudiante en ella, y tiene asignada un peso establecido a priori por el profesor, de tal forma que la posibilidad de que una sea elegida viene dada por la fórmula:

$$pos_i = \frac{PesoArea_i}{\sum_j PesoArea_i} \quad (3.15)$$

92 CAPÍTULO 3. EL DIAGNÓSTICO EN LOS SISTEMAS TUTORES INTELIGENTES

Una vez seleccionada el área, se elige el ítem más informativo de entre los disponibles para ella. El nivel de información depende de la calificación actual del examinando y de la dificultad del ítem y se calcula de la siguiente forma:

$$inf_M = 1 - d_i \quad (3.16)$$

$$inf_I = \begin{cases} \frac{-4}{3}d_i + \frac{4}{3} & \text{si } d_i \geq 0,25 \\ 4d_i & \text{si } d_i < 0,25 \end{cases} \quad (3.17)$$

$$inf_S = \begin{cases} -2d_i + 2 & \text{si } d_i \geq 0,5 \\ 2d_i & \text{si } d_i < 0,5 \end{cases} \quad (3.18)$$

$$inf_N = \begin{cases} -4d_i + 4 & \text{si } d_i \geq 0,75 \\ \frac{4}{3}d_i & \text{si } d_i < 0,75 \end{cases} \quad (3.19)$$

$$inf_B = d_i \quad (3.20)$$

Estas relaciones se basan en un principio según el cual la información aumenta con la dificultad de forma relativa a la calificación del alumno. En cuanto al mecanismo empleado para estimar el conocimiento, se computa la probabilidad a posteriori, dados los resultados obtenidos en un ítem, utilizando la fórmula:

$$P(nota|result) = \frac{P(notaAnterior)P(result|nota)}{\sum_{nota} P(notaAnterior)P(result|nota)} \quad (3.21)$$

donde $P(notaAnterior)$ es la estimación del conocimiento del alumno obtenida anteriormente, $nota$ toma los valores M, I, S, N o B y $P(result|nota)$ representa la curva característica del ítem definida con la fórmula:

$$P(result|nota) = g_i + (1 - g_i)P_{ok} \quad (3.22)$$

donde g_i , aunque los autores no lo denominan así, parece representar el factor de adivinanza del ítem y,

$$P_{ok}(M) = 1 - d_i \quad (3.23)$$

$$P_{ok}(I) = \begin{cases} \frac{-4}{3}d_i + \frac{4}{3} & \text{si } d_i \geq 0,25 \\ 1 & \text{si } d_i < 0,25 \end{cases} \quad (3.24)$$

$$P_{ok}(S) = \begin{cases} -2d_i + 2 & \text{si } d_i \geq 0,5 \\ 1 & \text{si } d_i < 0,5 \end{cases} \quad (3.25)$$

$$P_{ok}(N) = \begin{cases} -4d_i + 4 & \text{si } d_i \geq 0,75 \\ 1 & \text{si } d_i < 0,75 \end{cases} \quad (3.26)$$

$$P_{ok}(B) = 1 \quad (3.27)$$

La decisión de finalizar el test se toma utilizando como criterio que el examinando haya respondido al menos al número mínimo de ítems prefijado por el profesor, y que la variación

sucesiva de las estimaciones se mantenga dentro de un umbral prefijado. Por último, para llevar a cabo la calibración, se utiliza el procedimiento que el utilizado en CBAT-2.

Este sistema ha sido probado con alumnos reales, pero las conclusiones a las que han llegado los autores no son muy esperanzadoras, ya que los resultados no ofrecen ninguna mejora sustancial con respecto a los TAI clásicos con evaluación bayesiana. Además justifican el uso de conjuntos difusos para la clasificación del alumno en que, en el peor de los casos, el funcionamiento es igual a un TAI bayesiano, y en que la elección de las relaciones difusas fue obtenida a partir de la experiencia propia de los profesores.

Por otro lado, el criterio de selección de ítems que se utiliza no aporta nada nuevo, puesto que si se observa con detenimiento, el proceso equivale a realizar subtests adaptativos de las áreas que se van a evaluar siguiendo un orden decreciente según su peso.

ALICE

ALICE (en inglés, *Adaptive Link Insertion in Concept-based Educational System*) (Kavcic et al., 2002) es SEAW diseñado en principio para cualquier dominio, que se adapta a las características de cada usuario. Tiene un modelo del dominio basado en una red de conceptos (grafo acíclico) vinculados con relaciones de prerrequisitos de dos tipos: *esencial*, si es crucial que el usuario sepa el concepto precedente, y *de soporte*, si saber el concepto precedente no es fundamental, pero sí recomendable. El modelo del alumno es de superposición sobre el dominio, en el que el conocimiento de un concepto se describe mediante una tripleta de funciones miembro para conjuntos difusos de conceptos desconocidos, conocidos y aprendidos. Este modelo se actualiza mediante la aplicación de reglas difusas.

La primera vez que el estudiante accede al sistema, se le propone un breve pretest para inicializar su modelo de usuario. Tras la inicialización, cada concepto estará completamente aprendido o bien requerirá más conocimientos. El modelo del alumno se irá actualizando en función de los resultados obtenidos en los tests convencionales, que va realizando durante el proceso de instrucción o sobre las unidades conceptuales visitadas. Si realiza un test de forma satisfactoria, se incrementan en su modelo los valores que representan cuánto sabe sobre los conceptos involucrados en el test. En caso contrario, no hay actualización del modelo. También se actualiza información de los conceptos visitados a través del material docente. Además, todos los conceptos relacionados con uno o más, evaluados mediante una relación de precedencia esencial, se actualizan aplicando propagación.

Este sistema combina el uso de lógica difusa con tests convencionales. El procedimiento de evaluación que se sigue se basa en la aplicación de tests de longitud fija, aunque los autores no ponen de manifiesto qué criterios se siguen para determinar la calificación del estudiante. Asimismo, tampoco explican cómo se actualizan su modelo del alumno en función de los resultados del test.

Evaluación difusa en INSPIRE

En el sistema INSPIRE, además de la implementación de PASS, se ha implementado otro módulo de diagnóstico (Grigoriadou et al., 2002) basado en una combinación entre lógica difusa y una aproximación a la decisión multicriterio. En este caso, para conocer el conocimiento del alumno en uno o más conceptos se utilizan tests de evaluación. Los ítems se agrupan utilizando la misma categorización correspondiente a las habilidades que el estudiante debe desarrollar: nivel de recuerdo, nivel de uso y nivel de búsqueda. También

94 CAPÍTULO 3. EL DIAGNÓSTICO EN LOS SISTEMAS TUTORES INTELIGENTES

se hace uso de un modelo cualitativo para clasificar al alumno, basado en cuatro habilidades. Cada pregunta será más útil (desde el punto de vista de la evaluación) para evaluar una cierta habilidad. Para reflejar esta característica, el profesor debe ponderar inicialmente esa importancia. Para ello, por cada pregunta y concepto que ésta evalúa, se define una matriz de 3×3 en la que se relacionan unos criterios con otros. Los valores de esta matriz α_{ij} reflejan la importancia de un criterio con respecto a otro en un rango de 1 (i es igual de importante que j) a 9 (i es mucho más importante que j). De esta forma, el peso de cada criterio se calcula utilizando la siguiente fórmula:

$$w_i = \frac{(\prod_{j=1}^3 \alpha_{ij})^{1/3}}{\sum_{i=1}^3 (\prod_{j=1}^3 \alpha_{ij})^{1/3}} \quad (3.28)$$

Los tests se evalúan utilizando un criterio porcentual. En función del porcentaje obtenido por el alumno, éste es clasificado en un nivel de habilidad. Se utiliza una función $f : U \rightarrow [0, 1]$, donde U puede tomar los valores 0, 10, 20, 30, 40, 50, 60, 70, 80 o 90, tal que si $f(10) = 0,6$, indica que si el alumno ha acertando un 10 % de preguntas, eso implica que pertenece al conjunto con un grado de 0,6. Asimismo se definirá una función, un conjunto difuso, por cada combinación de criterio (*Recuerdo, Uso, Búsqueda*) y habilidad (*insuficiente, casi insuficiente, casi suficiente, suficiente*), habiendo, por lo tanto, doce. La estimación del conocimiento se lleva a cabo de la siguiente forma: después de que el alumno responda a una pregunta, primero se divide el número de respuestas correctas en cada categoría por el número total de preguntas de la categoría, calculando de esta forma, el porcentaje de preguntas correctamente respondidas por categoría. Cada valor obtenido se redondea a la categoría difusa más cercana. Se obtendrían por lo tanto los porcentajes para las tres categorías x_R, x_U, x_B . A partir de estos valores, y tomando las doce funciones difusas, se calcula la siguiente matriz D:

$$\mathbf{D} = \begin{pmatrix} f_R^I(x_R) & f_R^{CI}(x_R) & f_R^{CS}(x_R) & f_R^S(x_R) \\ f_U^I(x_U) & f_U^{CI}(x_U) & f_U^{CS}(x_U) & f_U^S(x_U) \\ f_B^I(x_B) & f_B^{CI}(x_B) & f_B^{CS}(x_B) & f_B^S(x_B) \end{pmatrix}$$

A continuación se construye un vector con los pesos de la pregunta para cada una de las categorías, con respecto al concepto que está siendo evaluado $W = [w_R, w_U, w_B]$. Este vector se multiplica por la matriz D , y el resultado es otro vector P que representa el grado de pertenencia del conocimiento del alumno a los cuatro niveles de evaluación. La estimación final se calcula aplicando el *método del Centro de Gravedad*, según la siguiente fórmula:

$$v = \frac{\sum_{i=1}^4 ip_i}{\sum_{i=1}^4 p_i} \quad (3.29)$$

El valor resultante se redondea y a partir de él se obtiene el nivel de evaluación correspondiente. La selección de ítems se lleva a cabo siguiendo un criterio de dificultad incremental por categorías. Primero se muestra la pregunta más fácil de la categoría de recuerdo, luego de la de uso, posteriormente la de búsqueda y así sucesivamente. El criterio de finalización, parece que es, o bien el número máximo de preguntas disponibles o bien la propia iniciativa del examinando de interrumpir el test.

Entorno de tutorización personalizada, evaluación mediante tests y diagnóstico

Este STI (Hwang, 2003) sobre Internet se compone tres aplicaciones: 1) *METSAS*, una herramienta de autor que permite a los profesores añadir material educativo y construir bancos de ítems. 2) *PTS*, para que los profesores analicen los perfiles de alumnos utilizando técnicas de minería de datos. 3) *ITED*, que se compone de tres módulos: uno para la generación de tests; un módulo de diagnóstico para analizar los problemas de los estudiantes durante el aprendizaje, mediante la aplicación de métodos de diagnóstico basados en lógica difusa; y un módulo para guiar y supervisar el aprendizaje.

Este sistema genera tests de contenido balanceado. Para ello, cada ítem tiene asignado un vector que indica su importancia sobre los n conceptos (c_1, \dots, c_n). La relevancia e_{ij} de cada ítem Q_i sobre cada concepto C_j se mide en valores discretos entre 0 (no tiene relación con el concepto) y 5 (es muy importante para el concepto). La herramienta ITED se encarga de generar tests en dos pasos: primero calcula la dificultad de los ítems, y luego construye el test aplicando clusterización y programación dinámica.

1. La dificultad se calcula mediante lógica difusa. Para ello se utiliza como dato de entrada una combinación de los resultados obtenidos por alumnos que han realizado (con anterioridad) tests con esos ítems, y sus coeficientes de inteligencia. Tanto los resultados en los tests como los coeficientes de inteligencia se miden en tres niveles: bajo, medio o alto. Para decidir qué ítem debe formar parte de un test, se aplica un proceso de clusterización, utilizando una versión difusa de la red neuronal ART de dos capas (Carpenter y Grossberg, 1988), la cual recibe como entrada el vector de importancia de un ítem sobre cada concepto. La salida de la red es una agrupación de ítems en clases. En una de éstas estarán aquéllos que evalúen conceptos similares.
2. Para construir el test de contenido balanceado se utilizan técnicas de programación dinámica, a partir de la clasificación realizada por la red neuronal, la tabla de pesos de cada ítem sobre cada concepto, y la dificultad y discriminación de cada ítem.

Cuando un profesor quiere administrar un test tiene que indicar su dificultad y discriminación global. A partir de esa información, y utilizando las técnicas anteriormente descritas, se generará un test de contenido balanceado y con los parámetros requeridos. Es necesario reseñar que este sistema genera todas las cuestiones que van a ser administradas antes de que dé comienzo el test, y no conforme se va estimando el conocimiento del alumno como en los TAI.

Como conclusión, el autor resalta que el sistema produce óptimos resultados cuando el número de ítems es reducido, pero para un número considerable de ellos, éstos no son tan buenos.

3.4.4. Modelos de evaluación basados en redes bayesianas

Una de las ventajas del uso de redes bayesianas es que permiten representar distribuciones de probabilidad conjuntas necesitando menos espacio para ser almacenadas que una representación tabular (VanLehn y Martin, 1998). Por ejemplo, una red bayesiana con 10 nodos binarios podría representar una distribución de probabilidad conjunta sobre $2^{10} = 1024$ estados.

En esta sección sólo se recogen aquellas propuestas basadas en redes bayesianas en las que en algún momento se utilizan tests (o únicamente ítems) para el diagnóstico del conocimiento del alumno.

ANDES

ANDES (Conati et al., 1997, 2002) es un STI para enseñar Física Clásica Newtoniana, especialmente destinado a alumnos universitarios de Academias Navales y de secundaria, de escuelas controladas por el departamento de Defensa de los EEUU. ANDES mantiene modelos de alumnos que se actualizan mediante redes bayesianas, que contienen 540 reglas, las cuales permiten resolver 120 problemas de mecánica. Antes de poder utilizar estas reglas es necesario estimar la probabilidad a priori de que un alumno sea capaz de aplicarlas. Para determinar estos valores se utiliza un pretest en el que cada regla es tratada como una subtarea distinta (VanLehn et al., 1998). De esta forma, se determina qué regla domina cada alumno. Contando el número de veces que cada una de ellas es conocida por un alumno, es posible estimar su probabilidad a priori en una cierta población. Con este fin, se desarrolló un test de 34 ítems de opción múltiple y de respuesta corta. Si el examinando respondía correctamente a la pregunta, se asumía que era capaz de aplicar la regla asociada.

El principal problema de esta aproximación para el cálculo de las probabilidades a priori es que se asume que una o dos preguntas de tests son suficientes para poder determinar si el alumno conoce una regla. Los autores no han tenido en cuenta la posibilidad de que haya examinandos que acierten los ítems de opción múltiple al azar.

En la siguiente sección se estudian los dos sistemas para el diagnóstico y modelado del alumno utilizados en ANDES. Uno de los objetivos principales era substituir las técnicas heurística (carentes de garantías), utilizadas en la mayoría de los sistemas de modelado del alumno, por técnicas con un fundamento teórico subyacente.

OLAE y POLA

OLAE (del inglés, *On-Line/Off-line Assessment for Expertise*) (Martin y VanLehn, 1995; VanLehn y Martin, 1998) es un sistema para el modelado del alumno basado en la técnica de *seguimiento del conocimiento*. OLAE recoge datos de estudiantes mientras que resuelven problemas de introducción a la Física. A partir de esta información, es capaz de diagnosticar el conocimiento de éstos en 290 piezas de conocimiento (representadas mediante reglas). Utiliza las acciones del alumno para evaluar la probabilidad de que éste sepa las reglas físicas o algebraicas codificadas en su modelo de conocimiento. El conocimiento del estudiante sobre cada regla puede estar en tres estados:

- *No dominada*, si nunca aplicaría la regla.
- *Parcialmente dominada*, si el alumno aplicaría la regla pero sólo sobre papel y lápiz, y no mentalmente.
- *Dominada*, si aplica la regla, siempre que ésta sea aplicable.

La evaluación, y por tanto, la construcción del modelo del alumno es llevada a cabo mediante la propagación de evidencias, extraídas de sus acciones, en una red bayesiana construida sobre un grafo AND/OR. Este grafo, previamente construido por un experto, representa la solución del problema y está almacenado en una base de conocimiento. En él, cada regla es expresada mediante un *nodo regla* que puede estar en uno de los tres estados anteriores. Cada estado tiene una probabilidad que indica si el alumno está en él. Además, el grafo contiene tres tipos de nodos temporales:

- Los *nodos de aplicación* de una regla, pueden tener dos estados: aplicada o no aplicada, e indican si el sujeto ha aplicado la regla.
- Los *nodos proposición* pueden tener dos estados: en memoria o fuera de memoria, e indican si el individuo conoce la proposición.
- Los *nodos de acción* representan ecuaciones insertadas por el alumno durante la realización del ejercicio, indicando por tanto, si éste ha llevado a cabo la acción.

Los de aplicación pueden definirse como nodos AND, ya que todos sus antecedentes deben ser conocidos. Además cada uno de ellos se conecta con un *nodo hecho* que representa la conclusión que se deriva. Estos pueden ser definidos como nodos OR, porque cada conclusión puede alcanzarse de diversas formas.

Cuando el alumno termina de insertar sus soluciones, cada ecuación es normalizada y buscada en la base de conocimiento. Si se encuentra, representará una conclusión dentro del grafo; se añadirá, por tanto, como nodo acción del correspondiente nodo hecho, y se le asigna probabilidad 1. Luego, se propaga la evidencia. Cuando se lleva a cabo este procedimiento con todas las ecuaciones introducidas por el alumno, se realiza la predicción de su conocimiento.

OLAE permite modelar los siguientes tipos de actividades:

- a) *Actividades de resolución de problemas cuantitativos*, en los que el estudiante debe calcular parámetros físicos.
- b) *Ejemplos*, cuya interfaz es similar a las anteriores, y donde el alumno va viendo, de forma progresiva, cómo se resuelve un problema. El propio estudiante decide cuándo se le muestra cada paso de la resolución. Asimismo, el sistema infiere que el alumno conoce una regla aplicada en un paso, en función del tiempo que tarda en solicitar el siguiente.
- c) *Actividades de resolución de problemas cualitativos*, son ítems de opción múltiple.
- d) *Actividades de planificación de soluciones*, en las que el alumno tiene que explicar los pasos que deben seguirse en la resolución de un problema cuantitativo. Este tipo de ejercicios se resuelven analizando sintácticamente la respuesta dada, para comprobar si en ella aparecen palabras clave de las fases de resolución del problema.
- e) *Actividades de clasificación de problemas*, en las que el estudiante debe agrupar diferentes problemas en función de su similitud.

Para calibrar los parámetros de la red bayesiana se utiliza un procedimiento de EM, con el que se estiman los parámetros, tras aplicaciones sucesivas de OLAE a una muestra poblacional. La bondad de la calibración será tanto mejor cuanto mayor sea el tamaño de la muestra.

El problema principal que presenta OLAE es su ineficiencia. Por este motivo, la red bayesiana tuvo que ser drásticamente simplificada, y en vez de modelar 290 reglas, el conjunto fue restringido a tan sólo 25. Asimismo, muchas de las suposiciones que se realizaron en las actividades son un ciertamente discutibles. En el caso de los ejemplos, asumen que un estudiante conoce una regla por el tiempo que tarda en solicitar el siguiente. Esto es claramente una suposición que no se tiene por qué corresponder con la realidad. De la misma forma, la

corrección del resultado de las actividades de planificación de soluciones se basa únicamente en la identificación de palabras clave que deben aparecer en el test, lo cual también puede considerarse inadmisible.

POLA (en inglés, *Probabilistic On-Line Assessment*) (Conati y VanLehn, 1996; Conati et al., 2002) es un marco de trabajo para el modelado del alumno basado en una evaluación estadística de su comportamiento mientras que éste resuelve problemas. Hasta su aparición, el modelado del estudiante basado en probabilidades se fundamentaba en la realización de un *seguimiento de su conocimiento*. POLA fue desarrollado tomando como base OLAE. El objetivo inicial era la creación de un sistema que permitiera modelar al alumno aplicando las técnicas de *seguimiento del modelo* y de *seguimiento del conocimiento*.

En POLA se lleva a cabo una revisión de la red bayesiana de OLAE, reduciéndose el problema de ineficiencia en la propagación de evidencias. Esto es posible gracias a la construcción incremental de la red bayesiana, a partir del grafo de solución del ejercicio, y en función de las entradas proporcionadas por el estudiante. De igual forma, se solventaron problemas en OLAE referentes a la propagación no deseada de ciertas evidencias. Para ello, por cada acción, se determina qué reglas han sido utilizadas para derivarla. En caso de haber más de una, se consideran las distintas posibilidades aunque éstas se solapen. Estas derivaciones se introducen como un nuevo tipo de nodos: los *nodos de derivación*, que se enlazan con los de acción mediante un enlace XOR-débil, ya que se asume que el alumno raramente inferirá el mismo hecho dos veces. Asimismo, cada nodo de derivación es enlazado a los correspondientes nodos de aplicación (según el grafo del problema) mediante enlaces AND. Cada nodo de aplicación es enlazado con el nodo regla correspondiente mediante un AND-débil, ya que para una regla todos los antecedentes son conocidos, la regla será "*casi siempre*" aplicada.

Otros dos tipos de nodos introducidos para solventar los problemas de actualizaciones inadecuadas son los *nodos de solución* y los *nodos de redundancia*. Los primeros permiten representar de forma explícita los diferentes conjuntos de soluciones que puede tener un problema. Una vez que se identifican las posibles soluciones a un problema, POLA inicializa la red bayesiana del alumno añadiendo un nodo solución por cada conjunto de soluciones. Éstos son enlazados con un *nodo de redundancia*, que representa la probabilidad de que el alumno siga cada una de las soluciones o combinaciones de ellas.

Modelo Granularidad-Bayes de Tests Adaptativos

Es un modelo de evaluación que se inspira en CBAT-2 y que ha sido desarrollado por Collins et al. (1996), Collins (1996). Utiliza jerarquías de granularidad y redes bayesianas, ofreciendo un mecanismo de diagnóstico de múltiples rasgos u objetivos de aprendizaje en un único test. Las jerarquías permiten capturar los diferentes niveles de detalle en una red semántica, en la que se pueden representar relaciones de agregación y de prerrequisitos. La primera permite descomponer un concepto en subconceptos. Se definen también *clusters* para representar dos o más vistas del mismo concepto. Por ejemplo, para un dominio de aritmética básica (figura 3.3), un cluster podría estar formado por habilidades prácticas asociadas con la suma, resta, multiplicación y división, mientras que otro podría estar formado por los fundamentos teóricos de las cuatro operaciones. Los conceptos de un mismo *cluster* se relacionan entre sí mediante una relación AND, mientras que con los de otro cluster con una relación OR. Este último tipo de relación indica que es suficiente con uno de los componentes de la relación para constituir el conocimiento del concepto de nivel superior. Por último, dentro de esta jerarquía, los elementos finales (las hojas), llamados *observadores*, son

los ítems o incluso ejercicios, ejemplos, etc.; es decir, todo aquello que permite realizar un diagnóstico del conocimiento del alumno. Los padres de los observadores son las primitivas de los objetivos de aprendizaje, directamente medidas por los ítems.

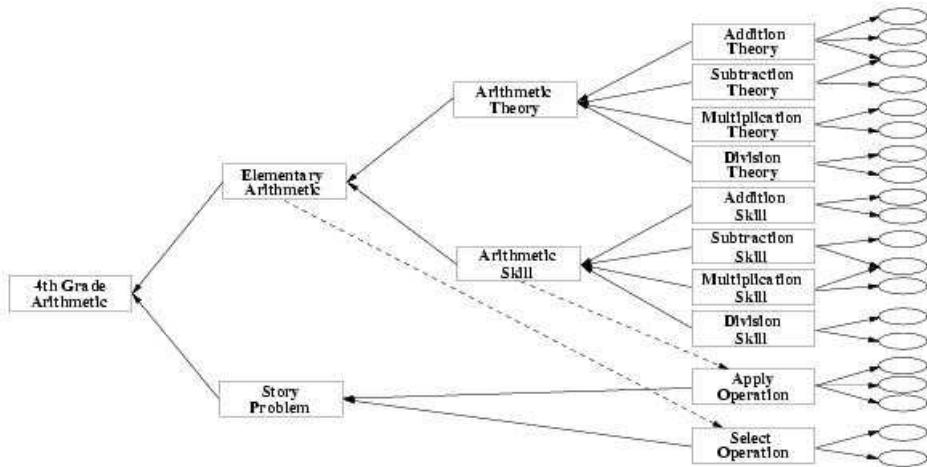


Figura 3.3: Estructura de la red bayesiana en el modelo Granularidad-Bayes (extraído de (Collins, 1996)).

Mediante el uso de jerarquías de granularidad aplicadas a los TAI, se intenta solventar el problema de garantizar tests de contenido balanceado. Además, la presencia de relaciones de prerequisitos contribuye a reducir el tamaño del test, ya que si un alumno ha demostrado que sabe un concepto, no se le harán preguntas sobre aquéllos que sean prerequisitos del concepto sabido. En esta propuesta pueden llevarse a cabo evaluaciones a cualquier nivel de granularidad de la jerarquía. Como la evaluación en un nodo de la jerarquía tiene efecto sobre otros, en virtud de la relación de agregación que se establece, se lleva a cabo una propagación del conocimiento. Para ello, sobre la jerarquía se superpone una red bayesiana que se construye de forma que los ítems dependen de los conceptos que evalúan. Como consecuencia, el profesor tiene que estimar dos probabilidades por ítem, algo que según Welch y Frick (1993) es empíricamente plausible. Si los ítems están asociados a más de un concepto, habrá que estimar 2^k probabilidades, donde k es el número de conceptos evaluados por el ítem.

Los tests finalizan cuando la probabilidad del nivel de conocimiento del alumno en el concepto (de mayor nivel en la jerarquía) que se está evaluando, es mayor que un cierto umbral especificado por el profesor, llamado *nivel de maestría*; o menor que otro umbral igualmente especificado por el profesor, llamado *nivel de no maestría*. Adicionalmente se permite relajar estos criterios de finalización cuando se deseé una evaluación más precisa en algunos de los nodos agregados. En este caso, no se finaliza el test hasta que todos los conceptos cuyo conocimiento se quiere estimar de forma precisa, hayan superado el umbral (de maestría o de no maestría) requerido. Como criterio de selección se utiliza el concepto de *utilidad* de un ítem i , definido por los autores de la siguiente forma:

$$\text{utilidad}_i \leftarrow (P(M_A|i) - P(M_A))(P(N_A) - P(N_A|\neg i)) \quad (3.30)$$

donde $P(M_A|i)$ es la probabilidad de saber el concepto A si el alumno responde correc-

tamente al ítem; $P(M_A)$ la probabilidad de saber el concepto A ; $P(N_A)$ la probabilidad de no saber el concepto A ; y por último, $P(N_A|\neg i)$ la probabilidad de no saber A si se responde incorrectamente al ítem i . Aplicando la ecuación 3.30, se eligen aquellos ítems que tienen mayor utilidad, que además van a coincidir con los de mayor discriminación, es decir, aquéllos que ofrecen más información para distinguir entre la maestría o no maestría.

Mediante esta propuesta, y gracias a la propagación bayesiana, se consigue la selección balanceada en los tests, y la determinación de las probabilidades que tiene el alumno de conocer los conceptos del currículo. En cuanto a cómo se decide si el test debe finalizar, los autores proponen dos criterios: el primero de ellos es que la probabilidad de saber el concepto raíz de la red bayesiana alcance un valor predeterminado; o por el contrario, que la probabilidad de no saber el concepto descienda por debajo de otro valor prefijado. El segundo criterio supone forzar que continúe la evaluación hasta que alguno de los dos umbrales de probabilidad anteriores sea alcanzado por todos los conceptos evaluados.

La principal desventaja de esta propuesta es inherente al uso de las redes bayesianas: el tiempo de cómputo es demasiado alto. Los autores hablan de que tan sólo el tiempo para seleccionar un ítem es de una media de 19 segundos. Otra desventaja (también común a todos los sistemas de redes bayesianas) reside en la determinación de la estructura de la red bayesiana. Los autores proponen varias alternativas, pero no manifiestan claramente el porqué de su elección final. Igualmente, la interpretación de las relaciones entre los nodos condiciona en exceso el proceso de inferencia del conocimiento del alumno y el número de probabilidades que deben calcularse a priori. En este caso, las probabilidades condicionadas a priori de cada ítem i ($P(M_A|i)$ y $P(N_A|\neg i)$) se calcularon a partir de un test administrado de forma convencional a 6235 individuos, como la proporción de los que habían respondido correctamente al ítem sobre el número total de los que habían aprobado el test; y la proporción de alumnos que habían respondido incorrectamente sobre el total de los que había suspendido el test. Al resto de probabilidades a priori se le asignó directamente el valor 0,5. Por último, este sistema sólo permite determinar el conocimiento en dos niveles, y su aplicación a escalas mayores multiplicaría los ya de por sí altos costes computacionales.

Por último, en comparación con el sistema anterior, puede verse que la relación entre los ítems y los conceptos que éstos evalúan se interpreta de dos formas completamente opuestas.

Tests Adaptativos Bayesianos

Esta técnica combina las redes bayesianas con los tests adaptativos para el diagnóstico y modelado del alumno (Millán, 2000; Millán et al., 2000; Millán y Pérez de la Cruz, 2002). Se trata de un modelo de superposición sobre un modelo del dominio basado en una jerarquía de granularidad con relaciones de agregación. Como se puede apreciar en la figura 3.4, esta jerarquía está formada por conceptos (partes minimales de conocimiento), temas (conjuntos de conceptos ponderados según su importancia relativa dentro del tema), y asignaturas (conjuntos de temas ponderados según su importancia relativa dentro de la asignatura). Estos elementos de la jerarquía forman parte de la red bayesiana en la que los nodos representan medidas del conocimiento del alumno. Se añaden también a la red bayesiana otro tipo de nodos, *nodos de recolección de evidencias*, que proporcionan información sobre el estado de conocimiento del alumno. Estos nodos corresponden a los ítems (en este caso, dicotómicos de opción múltiple) definidos en un TAI.

Para determinar la estructura de la red bayesiana de este modelo, se llevó a cabo un estudio exhaustivo en busca de la estructura que mejor se adecuaba a los objetivos marcados. La figura 3.4 muestra la red bayesiana utilizada. Los parámetros iniciales de la red se

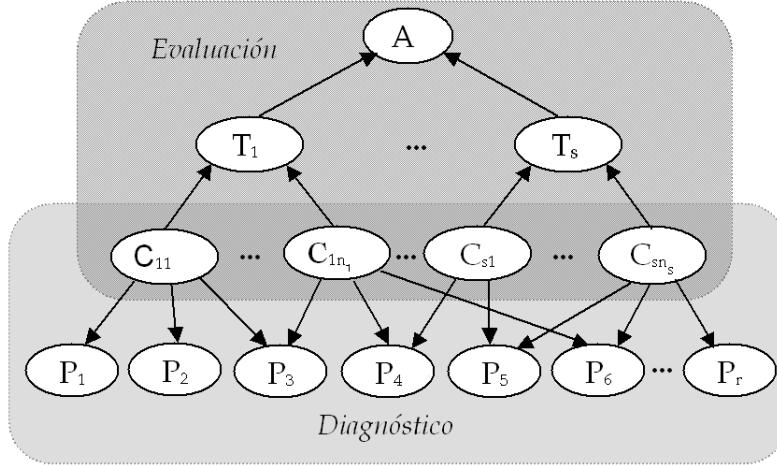


Figura 3.4: Estructura de la red para tests adaptativos bayesianos (extraído de (Millán, 2000)).

calculan de la siguiente forma: Si suponemos inicialmente que la probabilidad de que el alumno conozca un concepto i es un valor conocido $p(C = 1)$, la probabilidad de que conozca un tema T dado un conjunto de conceptos C_1, c_2, \dots, C_n evaluados en ese tema es igual a:

$$p(T = 1 | C_1, c_2, \dots, C_n) = \sum_{i=1, \dots, n} w_i \quad \forall i, C_i = 1 \quad (3.31)$$

donde w_i representa el peso del concepto C_i en el tema. La probabilidad de que el estudiante conozca el tema se calcula como sigue:

$$p(T = 1) = \sum_{i=1, \dots, n} w_i p(C_i = 1) \quad (3.32)$$

Análogamente la probabilidad de que el examinando sepa la asignatura A , condicionada porque su conocimiento en los temas T_1, \dots, T_m sea conocido, y la probabilidad de conocer la asignatura se calculan como indica la siguiente ecuación:

$$p(A = 1 | T_1, T_2, \dots, T_m) = \sum_{j=1, \dots, m} w_j \quad \forall j, T_j = 1 \quad (3.33)$$

$$p(A = 1) = \sum_{j=1, \dots, m} w_j p(T_j = 1) \quad (3.34)$$

donde w_j representa el peso del tema T_j en la asignatura.

Para el cálculo de las probabilidades de cada concepto $p(C_i)$ se utilizan TAI. Para ello, inicialmente, se lleva a cabo una transformación de la función logística 3PL, en virtud de la cual, el rango de la función pasa a ser de valores reales mayores o iguales que cero. La función que se utiliza como CCI viene expresada de la siguiente forma:

$$G(x) = 1 - \frac{(1 - c)(1 + e^{-1.7ab})}{1 + e^{1.7a(x-b)}} \quad (3.35)$$

Además, en este caso, los niveles de conocimiento admitidos son discretos, y representan el conocimiento del alumno en función de los conceptos que conoce, y son los siguientes: *no sabe nada, sabe un concepto, ..., sabe todos los conceptos menos 1, sabe todos los conceptos*. De esta forma, para N conceptos, el número de niveles de conocimiento es igual a $N + 1$. Como éstos son discretos, la CCI a su vez es también discretizada, pasando a ser un vector. Para el cálculo de los valores de ese vector, se asume que $G(0) = c$, donde c es el factor de adivinanza, cuyo valor es igual a 1 dividido entre el número de opciones de respuesta posibles; y que $G(N-1) = 1 - s$ donde s representa lo que se denomina el *factor de descuido*, es decir, la probabilidad de que un examinando conociendo un concepto, responda incorrectamente a una pregunta sobre éste. La probabilidad para el resto de niveles de conocimiento se computa por interpolación utilizando la ecuación 3.35.

En la aplicación que se hace de los TAI, es especialmente interesante destacar los criterios adaptativos de selección de ítems que se fundamentan, al igual que la aproximación de Collins et al. (1996), en el concepto de utilidad del ítem. En este caso, se ha definido la *utilidad* $U(P, C)$ de un ítem P para un determinado concepto C de dos formas distintas, en función del objetivo deseado: maximizar la información o priorizar los ítems de mayor sensibilidad y especificidad.

Es también interesante destacar el criterio de finalización en los tests con mecanismos de selección adaptativos. En este caso, el test finaliza si todos los conceptos han sido evaluados. Para determinar esto se fija un umbral s . Si la probabilidad de dominar un concepto es mayor o igual que $1 - s$, el concepto es sabido; por el contrario si es menor que s , es no sabido. Las probabilidades entre s y $1 - s$ indican que el conocimiento sobre el concepto no puede ser diagnosticado.

A diferencia de en la propuesta anterior, en este sistema no parece haberse estudiado la eficiencia del sistema en tiempo real, es decir, el tiempo requerido para realizar la selección del ítem en un test y para propagar los resultados a los nodos de la red. También hubiera sido útil estudiar cómo se comportan los criterios de selección en términos de número de ítems requerido, para determinar si merece la pena el uso de una red bayesiana o si por el contrario es mejor realizar un pequeño TAI de cada concepto evaluado. Hay que mencionar además, que se introducen ciertos componentes heurísticos en el cálculo de las probabilidades de los temas y la asignatura. El profesor debe establecer bajo su propio criterio unos pesos que determinan la relevancia de, por ejemplo un tema, para el cálculo de la probabilidad de saber la asignatura. Asimismo, no se menciona cómo se realiza la calibración de los ítems.

3.4.5. Conclusiones

En esta sección se ha llevado a cabo una revisión de las propuestas desarrolladas para el diagnóstico y modelado del alumno dentro del ámbito de los STI. Algunos de ellas son por sí mismos STI (ELM-ART, DCG, ActiveMath, TANGOW, INSPIRE, etc.); otras, o bien son (o podrían ser) módulos específicos de la arquitectura de un STI (PASS, Tests Adaptativos Bayesianos, etc.), o bien técnicas o algoritmos para el diagnóstico (CBAT-2, TAI basados en la T^a de la decisión, etc.). Todas han sido clasificadas en función de la técnica utilizada para llevar a cabo el diagnóstico del alumno. El grupo más grande se basan en el uso de heurísticos. Dentro de éstos, esta sección se ha centrado en los que usan tests, observándose que carecen de fundamento teórico. En general hacen uso de tests con un número fijo de

ítems (y algunos además con un tiempo límite para completarlo), en los que la evaluación se traduce en averiguar qué porcentaje de ítems han sido respondidos de forma correcta, o bien la suma de las puntuaciones obtenidas por el alumno en cada pregunta acertada. Sistemas como QUANTI y ELM-ART, intentan hacer uso de mecanismos de adaptación introduciendo tests ramificados, en los que una especie de árbol de decisión determina cuál es el siguiente ítem que debe ser mostrado al alumno en cada momento. Para ello, comparan la dificultad de los ítems con la estimación del conocimiento del alumno en ese momento. El problema de este tipo de aproximaciones es que el valor de la dificultad de los ítems es algo bastante subjetivo, que en la mayoría de los casos se ha obtenido según el criterio de un experto, y cuyo significado es algo que cada investigador interpreta de una forma diferente. Otra decisión que suelen tomar los expertos en estos sistemas es la longitud de los tests. La relevancia de este valor es con frecuencia menospreciada. Hay que tener en cuenta que si el número de ítems de un test es relativamente pequeño, esto puede provocar que el diagnóstico del conocimiento del alumno sea impreciso. Por el contrario, si este valor es excesivamente grande, puede provocar el rechazo del estudiante por aburrimiento.

Como conclusión sobre esta familia de sistemas, es necesario resaltar que el diagnóstico basado en heurísticos no parece adecuado dentro de STI en los que la importancia de un diagnóstico preciso es vital para garantizar una adecuada adaptación del proceso de instrucción.

El segundo grupo de sistemas son aquéllos que se basan en el uso de TAI. Dentro de esta familia hay dos tipos: por una parte, aquéllos que utilizan los modelos clásicos basados en TAI (estudiados en el capítulo anterior) como HEZINET o los desarrollados por Lilley y Barker (2004), Lilley et al. (2004). Estos sistemas suelen aplicar alguno de los modelo más utilizado en este ámbito, esto es, los modelos dicotómicos logísticos 1PL, 2PL o 3PL. Esta aproximación parece poco adecuada como mecanismo de diagnóstico en STI, ya que en los tests en los que se evalúan múltiples conceptos no existe ninguna garantía de que los ítems se seleccionen de entre todos los conceptos por igual. Asimismo, este tipo de tests multiconceptuales sólo son capaces de estimar el nivel de conocimiento global y no un diagnóstico detallado por concepto.

Otros sistemas basados en TAI intentan ofrecer soluciones para solventar el problema anterior. La gran mayoría de ellos se basan en el algoritmo CBAT-2 (*op. cit.*) o variantes, como las mencionadas Granularidad-Bayes, TAI borrosos, etc.; u otras más recientes como el sistema CALEAP-WEB (Gonçalves, 2004; Gonçalves et al., 2004). La razón principal está en que esta propuesta facilita la implementación de TAI, ya que ofrece un modelo logístico 3PL en el que los parámetros de las CCI o bien se determinan inicialmente y su valor no se modifica, o bien se calibran progresivamente durante la propia vida de sistema, sin necesidad de requerir una fase previa de calibración. Además, este algoritmo parece resolver otro gran problema de la aplicación de los TAI a los STI, la imposibilidad de evaluar simultáneamente y de forma balanceada diversas áreas de contenido en un mismo test. Esta segunda ventaja es parcial, ya que para obtener tests de contenido balanceado se recurre al uso de heurísticos, algo que parece que va en contra de la propia filosofía de los TAI, sistemas con un sólido fundamento teórico. En esta tesis se propone un modelo de diagnóstico basado en TAI pero que permite la evaluación simultánea de diversos conceptos en tests de contenido balanceado, sin necesidad de recurrir a heurísticos.

Otros sistemas se fundamentan en el uso de la lógica difusa (ALICE, INSPIRE, etc.) para inferir el conocimiento del alumno. El objetivo es formalizar, en algún sentido, los mecanismos de evaluación que utilizan los propios profesores. Para ello, combinan el uso de tests basados en heurísticos (o incluso TAI) con la evaluación utilizando conjuntos difusos

como medio de representación de la incertidumbre. En general, como se ha podido apreciar, no aportan ninguna novedad interesante desde el punto de vista del diagnóstico del alumno.

Por último, dentro de este análisis, se han estudiado las redes bayesianas. Este formalismo, basado en el teorema de Bayes, permite construir sistemas de diagnóstico bien fundamentados. De entre de este grupo, podemos destacar por afinidad con los TAI, el sistema de Tests Adaptativos Bayesianos y el modelo Granularidad-Bayes de Tests Adaptativos, que combina el uso de TAI como mecanismo de diagnóstico seguido de una propagación de estas evidencias en el modelo del alumno. Estos sistemas también combinan TAI y redes bayesianas con heurísticos similares a los utilizados en CBAT-2 para asegurar tests de contenido balanceado.

Es necesario hacer especial hincapié en que el uso de mecanismo de evaluación basados en fundamentos teóricos no garantiza por sí mismo que los resultados obtenidos sean correctos. Es necesario asegurarse de que las técnicas que se utilizan sí lo sean y de que no se hagan suposiciones erróneas. Este aspecto fue puesto de manifiesto por VanLehn y Martin (1998) con el uso de redes bayesianas. Ellos apuntaron que es preciso asegurar de alguna forma que en la construcción de la red bayesiana no se estén llevando a cabo suposiciones erróneas. En general, el diseño de la red bayesiana, y por tanto, la interpretación que se dé a las relaciones que ésta representa condicionarán el procedimiento y los resultados del diagnóstico del conocimiento del alumno. Por último, otro problema que tienen las redes bayesianas, es que suelen ser lentas. Este aspecto restringe bastante su uso en sistemas de diagnósticos dinámicos como los tests adaptativos, en los que el siguiente ítem del test se calcula en tiempo real, requiriendo por tanto procedimientos computacionalmente eficientes.

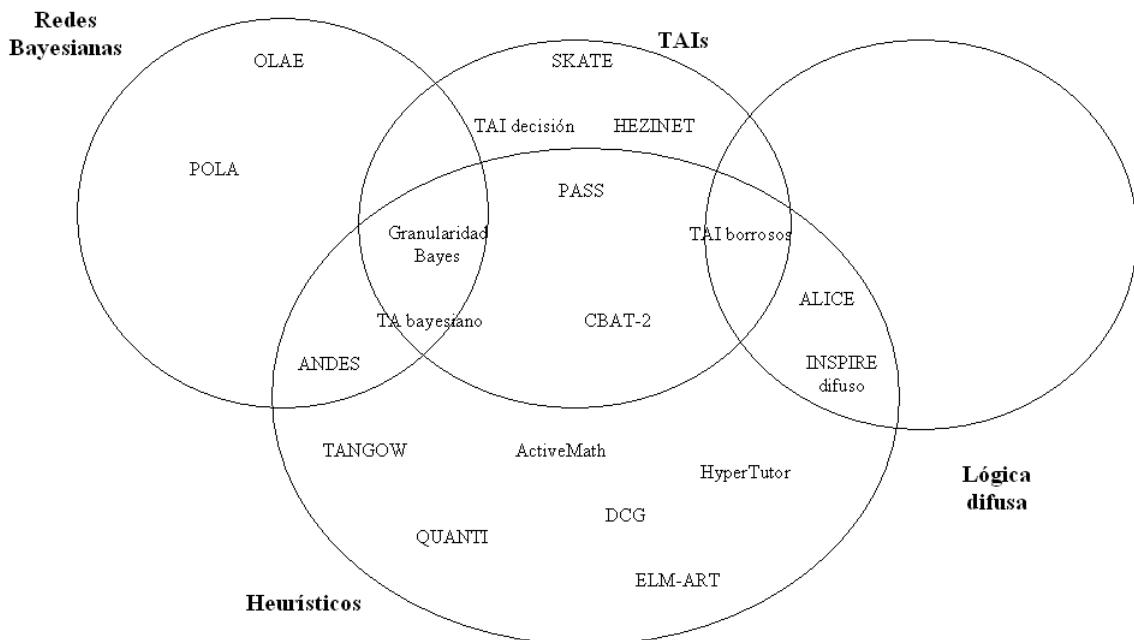


Figura 3.5: Clasificación de los sistemas presentados según su paradigma principal.

En la figura 3.5 se han representado los sistemas estudiados categorizados según la técnica de diagnóstico que utilizan. Como se puede apreciar, la mayoría de los sistemas

hacen uso de heurísticos en alguna de sus fases. Por ejemplo, los sistemas de TAI suelen utilizarlos para la generación de tests de contenido balanceado. De hecho, los sistemas como HEZINET que aparecen como puramente adaptativos, suelen limitarse a medir un único rasgo latente, y no ofrecen información sobre si generan tests de contenido balanceado.

	ELM-ART	DCG	ActiveMath	TANGOW	HyperTutor	QUANTI
<i>Modelo dominio</i>	red semántica	red semántica	red semántica	red semántica	red semántica	red semántica
<i>Modelo alumno</i>	superposición, inspeccionable de 4 capas	superposición	superposición +preferencias +aptitudes	superposición +preferencias	perfíles de usuario	superposición +perfíles de usuario
<i>Mecanismo evaluación</i>	tests adaptativos heurísticos	tests convencionales	tests convencionales	tests convencionales	tests convencionales	tests ramificados
<i>Tipos de ítems</i>	V/F, OM, RM, RC, completar	OM, RC, ordenación	OM + herramientas externas		OM	OM
<i>Tipos evaluación</i>	tests introductorios, tests finales, ejercicios	pretest y tests durante instrucción	tests durante instrucción	tests durante instrucción	tests durante instrucción	pretests
<i>Inicialización mod. alumno</i>	tests introductorios (máx. 25 ítems)	pretest	manual por parte del alumno	pretest		pretest
<i>Actualización mod. alumno</i>	ejercicios y tests finales	tests	actualizadores incremental y bayesiano	tests		
<i>Criterio selección</i>	adaptación heurística	no adaptativo	no adaptativo	no adaptativo	heurístico	ramificado en árbol
<i>Contenido balanceado</i>	tests en grupos de ítems	tests por concepto	tests por concepto	tests por concepto	tests por concepto	tests por concepto
<i>Criterio evaluación</i>	heurístico	heurístico	heurístico	heurístico	porcentual	por puntos
<i>Criterio finalización</i>	factor de confianza o número de ítems	número de ítems	número de ítems	número de ítems	número de ítems	número de ítems

Tabla 3.1: Características generales de los sistemas de evaluación basados en heurísticos.

Leyenda: Las siglas utilizadas son: V/F (ítems verdadero/falso), OM (de opción múltiple), RM (de respuesta múltiple) y RC (de respuesta corta).

En las tablas 3.1, 3.2 y 3.3 se han resumido algunas de las características principales de la mayoría de los sistemas descritos en esta sección. Las celdas en blanco indican que la característica correspondiente no está especificada en la bibliografía. El sistema SKATE no se ha incluido porque no se dispone de información suficiente sobre él. Los sistemas ANDES, OLAE y POLA tampoco se han incluido porque no utilizan tests para el diagnóstico.

En las tres tablas, las columnas representan las propuestas estudiadas a lo largo de esta sección. Las filas representan sus características: las dos primeras identifican cómo se han implementado los modelos del dominio y del alumno; la tercera indica el paradigma de evaluación utilizado; la cuarta recopila los tipos de ítems incluidos en el sistema correspondiente; la fila "tipos evaluación" contiene los diferentes tipos de pruebas de evaluación que se pueden identificar; las sexta y séptima indican cómo se inicializa y actualiza el modelo

	CBAT-2	HEZINET	Modelo T ^a decisión	INSPIRE	PASS
<i>Modelo dominio</i>	red semántica			red semántica	red semántica
<i>Modelo alumno</i>	superposición			superposición	superposición
<i>Mecanismo evaluación</i>	TAI	TAI	TAI basado en T ^a decisión	TAI	TAI
<i>Tipos de ítems</i>					
<i>Tipos evaluación</i>		tests durante instrucción		tests durante instrucción	pretests, tests durante instrucción
<i>Inicialización mod. alumno</i>	TAI	TAI		nivel de conocimiento medio	pretests
<i>Actualización mod. alumno</i>	TAI	tests de sesión, de curso y de admisión		tests de autoevaluación, de recapitulación	pretests, tests autoevaluación, recapitulación
<i>Criterio selección</i>	máxima información + heurísticos	basado en TRI	adaptativos basados en la T ^a decisión	máxima verosimilitud	máxima verosimilitud
<i>Contenido balanceado</i>	balanceo heurístico			balanceo heurístico	
<i>Criterio evaluación</i>	bayesiano	basado en TRI	bayesiano	máxima verosimilitud	
<i>Criterio finalización</i>	umbral de precisión o número de ítems			umbral de precisión o número de ítems	

Tabla 3.2: Características generales de los sistemas de evaluación basados en TAI.

Leyenda: Las siglas utilizadas son: V/F (ítems verdadero/falso), OM (de opción múltiple), RM (de respuesta múltiple) y RC (de respuesta corta).

del alumno respectivamente; las filas octava, décima y undécima señalan los criterios de selección de ítems, de evaluación y de finalización del test, respectivamente; por último, la novena indica cómo estos sistemas aseguran que los tests sean de contenido balanceado.

3.5. Discusión y conclusiones generales del capítulo

En este capítulo, se han puesto de manifiesto algunos de los problemas propios del diagnóstico en STI. Por un lado la incertidumbre inherente al propio proceso de diagnóstico, y por otra parte, el problema de la determinación inicial del estado de conocimiento del alumno, previo a cualquier interacción con éste. En principio, los TAI podrían utilizarse como mecanismo para el diagnóstico del conocimiento del alumno en STI. Huang (1996a) señala dos principales desventajas en el uso de este tipo de tests, desde el punto de vista de los STI: (1) La investigación en tests adaptativos está principalmente orientada a la aplicación de tests estándar a gran escala, diseñados por centros como el *Educational Testing Service*. (2) El proceso de calibración de los ítems de un TAI alarga enormemente el tiempo

transcurrido entre el inicio del desarrollo de un TAI y su aplicación definitiva, y constituye el principal freno con que cuenta su progreso en la actualidad. Los algoritmos desarrollados requieren extensos estudios empíricos para calibrar el banco de ítems (Wainer, 1990), que son difícilmente abordables por organizaciones pequeñas que ofrecen evaluaciones en línea a sus alumnos (o empleados). Esta deficiencia convierte en prohibitivo el uso de algoritmos de tests adaptativos en entornos de aprendizaje computerizados. Debido a esta gran inversión inicial, muchos proyectos e iniciativas dirigidas hacia la construcción tests de este tipo se han visto frustradas, y como consecuencia de ello se ha llegado a la paradójica situación de que en la actualidad, exceptuando algunos pocos países como los EEUU, Holanda o Israel, el número de TAI comercializados, y por tanto su uso, es considerablemente inferior a lo esperado, si se atiende a sus ventajas tantas veces elogiadas (Muñiz, 1997).

Otro problema de la aplicación de tests adaptativos es el problema del balanceo de contenido. Muchos algoritmos basados en tests adaptativos son de contenido oculto, es decir, la estrategia que utilizan para seleccionar ítems no tienen en cuenta el área de contenido dentro del currículo que la pregunta evalúa: no son capaces de realizar una selección balanceada. Por otro lado, la inferencia del nivel de conocimiento del alumno, así como el mecanismo de selección de ítems a través de la TRI, requieren un alto coste computacional. Esto es debido a que ambos procesos trabajan sobre un dominio continuo de los números reales.

El objetivo principal de esta tesis, será el desarrollo de un modelo de diagnóstico, en el que, sin perder el rigor teórico, se intenten superar este conjunto de trabas que dificultan el uso de TAI en los STI.

	TA conjuntos borrosos	ALICE	INSPIRE difuso	Entorno tutorización personalizada	Granularidad- Bayes	TA Bayesianos
<i>Modelo dominio</i>	red semántica	red semántica	red semántica	red semántica	red semántica	red semántica
<i>Modelo alumno</i>	superposición	superposición	superposición		superposición	superposición
<i>Mecanismo evaluación</i>	TAI	tests convencionales + lógica difusa	tests basados en lógica difusa y decisión multicriterio	tests convencionales	TAI	TAI
<i>Tipos de ítems</i>						OM, ejercicios
<i>Tipos evaluación</i>	pretests, tests durante instrucción	pretests, tests durante instrucción		tests durante instrucción		
<i>Inicialización mod. alumno</i>	pretest	pretest			nivel de conocimiento medio	
<i>Actualización mod. alumno</i>				tests convencionales		
<i>Criterio selección</i>	heurísticos	no adaptativa	dificultad incremental	ítems seleccionados antes de test mediante red neuronal	máxima información + heurísticos	basados en la utilidad y la máxima información
<i>Contenido balanceado</i>	balanceo heurístico			técnicas de programación dinámica	balanceo heurístico	balanceo heurístico
<i>Criterio evaluación</i>	similar a bayesiano		porcentual		bayesiano	bayesiano
<i>Criterio finalización</i>	umbral de precisión + número de ítems	número de ítems	número de ítems		umbral de precisión o número de ítems	umbral de precisión o número de ítems

Tabla 3.3: Características generales de los sistemas de evaluación basados en lógica difusa y redes bayesianas.

Leyenda: Las siglas utilizadas son: V/F (ítems verdadero/falso), OM (de opción múltiple), RM (de respuesta múltiple) y RC (de respuesta corta).

Parte III

PLANTEAMIENTO

Capítulo 4

Un modelo de respuesta discreto basado en la TRI

*La confianza, como el arte,
nunca viene de tener todas las respuestas,
sino de estar abierto a todas las preguntas.*
E.W. Stevens

En los dos capítulos anteriores, se han asentado las bases sobre las que se cimenta esta tesis. Por un lado, se ha puesto de manifiesto el problema del diagnóstico y de la inicialización del modelo del alumno en los STI. Sin un buen método de diagnóstico no se obtendrá un modelo preciso y fiable del alumno que permita al STI realizar una adecuada instrucción personalizada.

Por otra parte, en la sección 2.9 se enumeraron los elementos que intervienen en la generación de TAI. Así, para poder definir adecuadamente un modelo de diagnóstico basado en TAI, es requisito indispensable describir el modelo de respuesta que éste utilizará como fundamento teórico, y por tanto, como motor de inferencia adaptativa. Éste representa pues, el núcleo central de la teoría de los TAI, y se utilizará como parte del modelo de diagnóstico cognitivo que será descrito en el capítulo siguiente.

En la siguiente sección, se realiza una descripción de los tres aspectos que caracterizan al modelo presentado. Posteriormente, se procederá a definir formalmente las funciones de selección de una opción de respuesta, y de evaluación, así como las curvas característica de opción y de respuesta. En la sección 4.5, se presentan los distintos tipos de ítems que incluye el modelo, y se describe cómo se calculan sus curvas características. A continuación, y debido al carácter intratable en tiempo real (desde el punto de vista computacional) de algunos de esos ítems, se presenta una aproximación cuasipolítómica de este modelo de respuesta. Por último, se describe el algoritmo utilizado para calibrar los ítems.

4.1. Descripción general del modelo

El modelo de respuesta que se propone tiene tres características fundamentales: es discreto, no paramétrico y polítómico, cuya elección se justifica en esta sección.

4.1.1. Modelo discreto

La gran mayoría de los modelos de respuesta basados en la TRI son continuos, es decir, el nivel de conocimiento (o genéricamente, el rasgo latente) se suele medir utilizando como dominio los números reales. Este tipo de modelos son más precisos debido a la escala de medición que utilizan. A pesar de ello, desde el punto de vista puramente práctico, son más difíciles de utilizar. Algunas de las razones se enumeran a continuación:

- Se pueden hacer implementaciones computacionalmente más eficientes con modelos discretos que, como se verá en el transcurso de este capítulo, permiten utilizar soluciones polítómicas. Además, hacen posible abordar modelos multidimensionales de forma más eficiente que son prácticamente inabordables mediante aproximaciones continuas.
- El uso de modelos continuos en la teoría de los TAI requiere la aplicación de integrales que deben ser aproximadas, las cuales en ese proceso, terminan siendo discretizadas. Como resultado, los modelos continuos basados en la TRI estiman el nivel de conocimiento del alumno, y la probabilidad de certeza de esa estimación. Para ello, es necesario aplicar algoritmos iterativos de aproximación, costosos computacionalmente, tales como Newton-Raphson. Por el contrario, con un modelo discreto, como el presentado en este capítulo, el conocimiento del alumno se expresa mediante pares nivel de conocimiento/probabilidad asociada, que permiten inferir casi directamente cuál es el nivel de conocimiento.
- Por último, las escalas de evaluación que los profesores utilizan comúnmente son discretas. Es decir, los valores que se obtienen como resultado en el modelo continuo son finalmente discretizados. Aunque a veces se utilicen inicialmente valores continuos, finalmente lo que se hace es una clasificación del alumno en grupos según su dominio en la asignatura o concepto evaluado.

Por estas razones, se ha preferido utilizar un modelo discreto, que permite representar las curvas características y las estimaciones de conocimiento del alumno como vectores. Asimismo, entre las ventajas que aporta el uso de este tipo de modelos se pueden destacar las siguientes:

- El número de niveles de conocimiento en que se puede evaluar al alumno es variable. Esto permite que si el profesor desea llevar a cabo evaluaciones con una mayor granularidad sólo tenga que incrementar el número de niveles de conocimiento. Lo mismo ocurre en caso contrario, si por ejemplo el profesor sólo desea averiguar si los estudiantes son aptos en una determinada materia puede reducir el número de niveles a tan sólo dos.
- Permiten ver el problema del diagnóstico del conocimiento del alumno como un caso de clasificación, en el que se ha optado por utilizar la TRI, en vez de aplicar soluciones derivadas de las redes neuronales, algoritmos genéticos o la minería de datos.

Por el contrario, el uso de modelos discretos también conlleva ciertos inconvenientes:

- Cuando el nivel de conocimiento real de un alumno está entre dos niveles, con un TAI, el proceso de inferencia de este valor sufre un problema de estabilidad. Esto es debido a que el algoritmo correspondiente no es capaz de decantarse entre dos niveles

de conocimiento, y como consecuencia, cuando la selección del ítem es adaptativa, el número de preguntas necesario para diagnosticar el conocimiento del sujeto es mayor de lo habitual.

- Cuando se desea clasificar a un estudiante empleando un modelo continuo, las categorías resultantes pueden ser de cualquier tamaño; pero en los modelos discretos, la posterior división en categorías debe hacerse sobre la estructuración en niveles que establece el modelo. Por ejemplo, si se deseara determinar la aptitud de un individuo en dos estados (apto y no apto), y el número de niveles establecido a priori es tres, no es posible hacer una división en dos partes iguales.

4.1.2. Modelo no paramétrico

Como se puso de manifiesto en la sección 2.8, la mayor parte de los modelos de respuesta basados en la TRI son paramétricos. Su uso supone asumir que el patrón de respuesta de los alumnos, según su nivel de conocimiento, viene siempre descrito por una función conocida (Sijtsma y Molenaar, 2002). Esta suposición es aún más fuerte en el caso de modelos polítómicos, ya que se asume que todas las opciones de respuesta vienen descritas por funciones conocidas que deben ser estimadas a partir de un proceso de calibración. En la realidad, el ajuste de los datos reales a los modelos no siempre es tan preciso como debiera esperarse. Por este motivo, antes de calibrar paramétricamente las curvas características de los ítems, se suele llevar a cabo un análisis no paramétrico de éstas, mediante el análisis de los residuos o con χ^2 . Una vez realizada esta tarea, se estudia la similitud de los datos con el modelo paramétrico para ver si realmente se ajusta a él. En caso contrario, el ítem se descarta y por tanto se elimina. Desde el punto de vista de la estadística moderna y de la teoría de la medida, es inaceptable calibrar directamente un ítem paramétricamente sin estudiar si realmente los datos se ajustan al modelo (Stout, 2001).

En muchos casos, el problema del uso de modelos paramétricos radica en que las curvas características no pueden ser modeladas paramétricamente de forma correcta. Esta razón de peso ha contribuido a que muchos investigadores tiendan al uso de aproximaciones no paramétricas para representar las curvas características. Para definir un modelo de este tipo basta con establecer un conjunto mínimo de condiciones que debe cumplir (Junker y Sijtsma, 2001b): monotonicidad, independencia local y unidimensionalidad.

Los modelos paramétricos pueden ser considerados como una malla sobre la que se colocan los datos reales, para caracterizar las propiedades generales de las respuestas a los ítems permitiendo, de esta forma, la realización de predicciones a partir de ellas. Los parámetros del modelo, estimados a partir de los datos, muestran si éste se ajusta a ellos, pero tienen la desventaja de que no son capaces de expresar propiedades de aquéllos que se alejen del modelo. Por otro lado, los modelos no paramétricos representan un malla más flexible, que ayuda a detectar fácilmente irregularidades en los datos (Junker y Sijtsma, 2001b).

Diversos autores (Junker y Sijtsma, 2001b; Stout, 2001; Molenaar, 2001) han puesto de manifiesto las ventajas del uso de modelos de respuesta basados en aproximaciones no paramétricas. Entre ellas, Junker y Sijtsma (2001b) resumen y destacan las siguientes:

1. Permiten comprender mejor los modelos paramétricos y estudiar sus propiedades.
2. Son aplicables en ocasiones en las que los modelos paramétricos no se ajustan adecuadamente a los datos, haciendo que los resultados de la calibración no sean correctos. A través de la TRI no paramétrica es posible detectar estos casos.

3. Requieren una muestra poblacional de menor tamaño para calibrar los ítems que los modelos paramétricos. Según Molenaar (2001), hacen un uso más económico de los datos de entrada utilizados para la calibración.

Asimismo, Junker (2000) señala que, en virtud de estas características, los modelos no paramétricos basados en la TRI pueden ser de gran utilidad, cuando se aplican al diagnóstico cognitivo.

4.1.3. Modelo politómico

En capítulos anteriores se han descrito las ventajas del uso de este tipo de modelos en la estimación del conocimiento del alumno: inferencias más precisas, tests con un número menor de ítems, etc. Por otro lado, el principal inconveniente es que requieren una muestra poblacional considerablemente mayor que en el caso dicotómico para poder calibrar satisfactoriamente los ítems.

4.2. Definiciones formales

A continuación se introducirán las definiciones de *función de respuesta seleccionada* y de *evaluación de la respuesta a un ítem*, que serán útiles para comprender y formalizar el modelo. Después de esto, se explicará con detalle cómo se ha realizado su discretización, y posteriormente, se definirán dos tipos de curvas características, piezas esenciales para comprender su funcionamiento. Asimismo, se enumerarán los diferentes tipos de ítems para los que se ha concebido este modelo de respuesta. Finalmente, se abordará cómo se lleva a cabo la calibración de los ítems en este modelo.

Supóngase un determinado ítem i , generalmente, formado por un enunciado y un conjunto de posibles opciones de respuesta (o partes de la respuesta), que serán identificadas por r_{ij} , siendo j un valor que permite identificar esa opción de respuesta dentro del ítem i de forma unívoca. Consecuentemente, y asumiendo que el ítem tiene m_i posibles respuestas, el conjunto de opciones de respuesta vendría definido de la siguiente forma: $\{r_{i1}, r_{i2}, \dots, r_{im-1}, r_{im}\}$. Asimismo, es posible que el alumno no seleccione ninguna de las opciones, es decir, podría dejarlo en blanco. Para representar esta situación, se añadirá una nueva alternativa r_{i0} , a través de la cual se expresa que el sujeto no ha seleccionado ninguna opción. El conjunto M_i de opciones de respuestas para el ítem i queda finalmente de la siguiente forma: $M_i = \{r_{i0}, r_{i1}, r_{i2}, \dots, r_{im-1}, r_{im}\}$.

Definición 4.1 (Función respuesta seleccionada). Sea un ítem i , se define la función respuesta seleccionada del ítem $S_i : M_i \rightarrow \{0, 1\}$ de la siguiente forma:

$$S_i(r_j) = \begin{cases} 1 & \text{si el alumno ha seleccionado la opción } j \\ & \text{como respuesta al ítem } i \\ 0 & \text{en otro caso} \end{cases} \quad (4.1)$$

Definición 4.2 (Función de evaluación de la respuesta). Sea el ítem i , se define la función de evaluación de la respuesta del ítem, como una función $R_i : \{0, 1\} \times \dots \times \{0, 1\} \rightarrow \{0, 1\}$. Esta determina, si el patrón de opciones de respuesta, seleccionadas por el alumno para el ítem i , es correcto. Formalmente, esto se puede expresar:

Sea $\vec{u}_i = \{S_i(r_1), S_i(r_2), \dots, S_i(r_m)\}$ el patrón de opciones de respuesta seleccionadas por el examinando, la función de respuesta correcta del ítem se define de la siguiente forma:

$$R_i(\vec{u}_i) = \begin{cases} 1 & \text{si el patrón de respuestas es correcto} \\ 0 & \text{en otro caso} \end{cases} \quad (4.2)$$

Nótese que la respuesta de un alumno a un ítem i es un subconjunto de M_i , cuya cardinalidad oscilará entre 0 y la cardinalidad de M_i .

4.3. Discretización del modelo

La discretización consiste en modificar el rango de valores que puede tomar el nivel de conocimiento del alumno, reduciéndolo a un conjunto finito y ordenado de valores. En este caso, los valores permitidos son números naturales entre 0 y el número de niveles de conocimiento menos 1. Sea K el número de niveles de conocimiento, 0 representa la ausencia total de conocimiento, y $K - 1$ el conocimiento absoluto. La escala de valores posibles que puede tomar el nivel de conocimiento será, por tanto: 0, 1, 2, 3, ..., $K - 2$, $K - 1$. El número total de niveles de conocimiento K será definido por el profesor, según la granularidad que desee en la evaluación. La única restricción que se impone es que deberá ser un número natural mayor que uno.

Como consecuencia, las curvas características (que como se vio en la sección 2.8 modelan la probabilidad de que un alumno responda correctamente al ítem dado su nivel de conocimiento) se representan mediante vectores de probabilidades, $P_i(R_i(\vec{u}_i) = 1|\theta) : \{0, 1, \dots, K - 1\} \rightarrow [0, 1]$. Al tratarse de una discretización, la CCI puede verse como un vector, en el que cada valor representa la probabilidad de que el alumno responda correctamente al ítem i , dado su nivel conocimiento. Por tanto, la CCI de un determinado ítem vendrá descrita de la siguiente forma:

$$P(R_i(\vec{u}_i) = 1|\theta) = [P(R_i(\vec{u}_i) = 1|\theta = 0), P(R_i(\vec{u}_i) = 1|\theta = 1), P(R_i(\vec{u}_i) = 1|\theta = 2), \dots, P(R_i(\vec{u}_i) = 1|\theta = K - 2), P(R_i(\vec{u}_i) = 1|\theta = K - 1)]$$

A partir de esto, puede calcularse el vector correspondiente a la probabilidad de que el alumno responda incorrectamente a la pregunta, dado su nivel de conocimiento:

$$P_i(R_i(\vec{u}_i) = 0|\theta) = \overrightarrow{1} - P_i(R_i(\vec{u}_i) = 1|\theta)$$

Asimismo, las distribuciones de probabilidades del conocimiento del alumno también se representan mediante vectores:

$$P(\theta = 0|\vec{\mathbf{u}}), P(\theta = 1|\vec{\mathbf{u}}), \dots, P(\theta = K - 2|\vec{\mathbf{u}}), P(\theta = K - 1|\vec{\mathbf{u}})$$

Cada valor representa, en este caso, la probabilidad de que el nivel de conocimiento del examinando sea el correspondiente, dado $\vec{\mathbf{u}}$. El vector $\vec{\mathbf{u}} = \{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_n\}$, representa al conjunto de patrones de respuesta que el alumno ha seleccionado en los n ítems que le han sido administrados en el test.

4.4. Curvas características de respuesta y de opción

Por tratarse de un modelo politómico, por cada posible patrón de respuesta se define una curva característica denominada *Curva Característica de Respuesta*, de la siguiente forma:

Definición 4.3 (Curva Característica de Respuesta (CCR)). Esta función $P_{i\vec{u}_i}(\vec{u}_i|\theta) : \{0, 1, \dots, K - 1\} \rightarrow [0, 1]$ representa la probabilidad de que el alumno seleccione un determinado patrón de respuesta \vec{u}_i , dado su nivel de conocimiento. La CCR, al igual que la CCI, será un vector, en el que cada valor representa la probabilidad de que el alumno haya seleccionado ese patrón de respuestas, dado su nivel de conocimiento.

Los posibles patrones de respuesta para un determinado ítem conforman un espacio probabilístico donde los sucesos son precisamente los patrones. Como resultado, el espectro de respuestas potenciales deberá copar todo el espacio probabilístico. De esta forma, las CCR deberán cumplir siempre la siguiente condición:

Axioma 4.1. Sea W_i el conjunto de todos los posibles patrones de respuesta del ítem i :

$$\sum_{\vec{u}_i \in W_i} P_{i\vec{u}_i}(\vec{u}_i|\theta) = \vec{1}_K \quad (4.3)$$

donde $\vec{1}_K$ es una distribución de probabilidades constante e igual a uno. Es decir, un vector de dimensión igual a K (el número de niveles de conocimiento), y en el que todos los valores son igual a uno.

Por otro lado, se definen también las *Curvas Características de Opción*, de la siguiente forma:

Definición 4.4 (Curva Característica de Opción (CCO)). Como se ha mencionado anteriormente, los ítems se componen de un enunciado, y de un conjunto de posibles opciones (o partes) de respuesta. En este modelo, cada una de esas opciones de respuesta tiene asociada una curva característica, que representa la probabilidad de que el alumno seleccione esa opción dado su nivel de conocimiento: $P_{ij}(S(r_j) = 1|\theta) : \{0, 1, \dots, K - 1\} \rightarrow [0, 1]$.

Consecuentemente, para un ítem se definirán tantas CCO como posibles opciones de respuestas tenga, incluyendo, en los casos en que sea posible, la respuesta en blanco. Es decir, para un ítem i , el conjunto de CCO que tendrá asociadas será el siguiente:

$$P_{i0}(S(r_0) = 1|\theta), P_{i1}(S(r_1) = 1|\theta), P_{i2}(S(r_2) = 1|\theta), \dots, P_{im}(S(r_m) = 1|\theta)$$

Las CCO están, por tanto, asociadas a cada opción dentro de un ítem. A partir de éstas, y en función del tipo de ítem, se construye la CCR correspondiente al patrón de respuestas seleccionado por el examinando. Es precisamente la CCR la que se utilizará como curva identificativa de la respuesta del alumno dentro del modelo TAI. Las CCR, en este modelo, sustituirán a la CCI del ítem (o a su opuesta), en los procesos de estimación del conocimiento del alumno y selección del siguiente ítem; en resumen, en diversas fases del algoritmo adaptativo.

En la siguiente sección, se enumeran los diferentes tipos de ítems definidos en el modelo, explicando, en cada caso, cuáles son los posibles patrones de respuestas y cómo se calcula su CCR a partir de las CCO de las opciones.

4.5. Tipos de ítems y cálculo de las CCR

El modelo propuesto, como la mayoría de los políticos utilizados en TAI, fue concebido inicialmente para ítems de opción múltiple. Posteriormente, han sido añadidos otros tipos que lo dotan de mayor potencia y expresividad. Esto ha permitido demostrar que el modelo admite cualquier tipo de ítem que se pueda transformar en uno de opción múltiple.

A continuación se enumeran los tipos de ítems del modelo. Por cada tipo se describe el formato del ítem, haciendo especial hincapié en las opciones de respuesta que le acompañan, ya que ésta es la característica que los distingue. Cada opción de respuesta tendrá asociada su CCO, que debe ser calibrada adecuadamente. Asimismo se verá cómo se infieren las CCR, en función de cómo sea el patrón de respuesta del ítem, y cómo se calcula la CCI a partir de estas últimas.

4.5.1. Ítems verdadero/falso

En ellos se pregunta acerca de la certeza de una sentencia recogida en el enunciado. Únicamente se muestran dos posibles respuestas: verdadero (o correcto) y falso (o incorrecto). En este tipo de ítems se restringe la posibilidad de que el alumno deje el ítem en blanco. Forzosamente deberá seleccionar una de las opciones. Como consecuencia, éstas coinciden con los posibles patrones de respuesta al ítem. Esto es, sea r_1 la respuesta "verdadero", y r_2 la respuesta "falso", el patrón de respuesta se puede expresar de la siguiente forma: $\vec{u}_i = \{S_i(r_1), S_i(r_2)\}$. El cálculo de la CCR se hace siguiendo la función 4.4:

$$P_{iu_i}(\vec{u}_i|\theta) = \begin{cases} P_{i1}(S(r_1) = 1|\theta) & \text{si } \vec{u}_i = \{1, 0\} \\ P_{i2}(S(r_2) = 1|\theta) & \text{si } \vec{u}_i = \{0, 1\} \end{cases} \quad (4.4)$$

Como se puede apreciar, la CCR correspondiente va a coincidir con la curva característica de la opción de respuesta seleccionada. Asimismo, como el espacio probabilístico sólo contempla dos sucesos, la suma de ambos será igual a una distribución uniforme e igual a uno. Por este motivo, el cálculo de una CCO a partir de la otra es inmediato:

$$P_{i1}(S(r_1) = 1|\theta) = \overrightarrow{1} - P_{i1}(S(r_2) = 1|\theta) \quad (4.5)$$

Al haber un único patrón de respuesta, la CCI corresponde con la CCR del patrón de respuesta correcto, y por transitividad con la CCO de la opción de respuesta correcta. Es decir,

$$P_i(R_i(\vec{u}_i) = 1|\theta) = P_{i1}(S(r_1) = 1|\theta) \quad (4.6)$$

4.5.2. Ítems de opción múltiple

En este tipo de ítems se muestran dos o más opciones de respuestas, de entre las que el alumno debe seleccionar una única. También tiene la posibilidad de dejar la respuesta en blanco. En este caso, al igual que en el anterior, las CCR van a coincidir con las CCO.

Sea el ítem i con el conjunto de $m+1$ posibles opciones de respuesta $M = \{r_0, r_1, \dots, r_m\}$, y sea el patrón de respuesta \vec{u}_i del alumno, la CCR se calcula de la siguiente forma:

$$P_{i\vec{u}_i}(\vec{u}_i|\theta) = P_{ij}(S(r_j) = 1|\theta), \quad r_j \in M \quad (4.7)$$

Todas las CCO de un ítem de este tipo deben cumplir que su suma sea igual $\vec{1}_K$, siendo K el número de niveles de conocimiento. Esto puede expresarse matemáticamente de la siguiente forma:

$$\sum_{j=0}^m P_{ij}(S(r_j) = 1|\theta) = \vec{1}_K \quad (4.8)$$

En este caso, la CCI también coincide con la CCR del patrón de respuesta correcta, y por tanto con la CCO de la opción correcta. Obsérvese además, que los ítems verdadero/falso pueden considerarse un caso particular de este tipo.

4.5.3. Ítems de respuesta múltiple

Son similares a los de opción múltiple, pero en esta ocasión el examinando puede elegir más de una opción de respuesta simultáneamente. La respuesta correcta es por tanto, una cierta combinación de entre esas opciones. Este tipo de ítems permite una puntuación parcial ya que el alumno puede acertar sólo algunos de sus componentes correctos. En el modelo propuesto estos ítems pueden ser clasificados de dos formas:

Ítems con opciones independientes

Aquí, la corrección de una de las opciones no influye en la del resto. Este tipo de ítems pueden considerarse como una colección, un *testlet*, de tantas cuestiones verdadero/falso como opciones de respuesta haya, propuestas una a una, donde se pregunta si la opción correspondiente es correcta, dado el enunciado. El ítem podrá ser evaluado, de esta forma, como completamente correcto, incorrecto o parcialmente correcto, en función del número de opciones seleccionadas adecuadamente.

Un ejemplo de este tipo de ítems, sería el siguiente: *¿Cuáles de los siguientes animales son mamíferos? a)Perro b)Rana c)Tigre d)Ballena e)Tiburón.* Este ítem sería equivalente a la siguiente colección, formada por cinco ítems: (1) *¿Es el perro un animal mamífero?* a)Sí b)No. (2) *¿Es la rana un animal mamífero?* a)Sí b)No. (3) *¿Es el tigre un animal mamífero?* a)Sí b)No. (4) *¿Es la ballena un animal mamífero?* a)Sí b)No. (5) *¿Es el tiburón un animal mamífero?* a)Sí b)No.

La respuesta en blanco no viene modelada por medio de una CCO ya que, por definición, el no seleccionar ninguna respuesta equivale a asumir que todas las opciones son falsas.

El cálculo de la CCR, dado el patrón de respuesta \vec{u}_i , sería el siguiente:

$$P_{i\vec{u}_i}(\vec{u}_i|\theta) = \prod_{j=1}^m P_{ij}(S(r_j) = 1|\theta)^{S(r_j)} P_{ij}(S(r_j) = 0|\theta)^{1-S(r_j)} \quad (4.9)$$

Al considerarse como una colección de ítems verdadero/falso, cada opción por sí sola constituye un espacio probabilístico. Esto implica que las opciones de respuesta entre sí no están sujetas a ningún tipo de restricción, a diferencia de lo que sucede en los ítems anteriores.

En este caso, la CCI también equivale a la CCR del patrón de respuesta correcto:

$$P_i(R_i(\vec{u}_i) = 1|\theta) = P_{i\vec{u}_i}(\vec{u}_i|\theta) \quad (4.10)$$

Ítems con opciones dependientes

Estos ítems tienen el mismo formato que los anteriores, diferenciándose en la semántica. En este caso, la respuesta correcta es uno o más subconjuntos de opciones. De esta forma, para que la respuesta a un ítem de esta clase se considere como correcta, es necesario que el examinando seleccione todas las opciones que representan un conjunto correcto.

Por ejemplo, considérese el siguiente ítem: *Señala de entre los siguientes países, un conjunto de ellos que fueran aliados durante la Segunda Guerra Mundial: a) Reino Unido b) Japón c) Alemania d) URSS e) EEUU.* Si se analiza el enunciado, habría dos conjuntos correctos de respuesta. Por un lado el conjunto formado por las opciones *a*, *d* y *e*; y por otro, el grupo compuesto por las opciones *b* y *c*. Efectivamente, ambas combinaciones son buenas, aunque obviamente el alumno sólo podrá seleccionar una de ellas.

Como se puede apreciar, dado un número m de posibles opciones de respuesta, para un ítem de este tipo, hay 2^m posibles patrones, ya que cada opción puede estar o no seleccionada. Asimismo, se incluye también la respuesta en blanco, que modela cuando el alumno no elige nada. Es decir, para un ítem con tres opciones *a*, *b* o *c*, los posibles patrones de respuesta son: respuesta en blanco, $\{a\}$, $\{b\}$, $\{c\}$, $\{a, b\}$, $\{a, c\}$, $\{b, c\}$ y $\{a, b, c\}$. Por tanto, este tipo de ítems puede considerarse un ítem de opción múltiple en el que las opciones son las $2^m - 1$ posibles combinaciones, exceptuando la respuesta en blanco. Por este motivo, el tratamiento de las CCO es especial, puesto que están asociadas a los conjuntos de opciones, y no a cada opción individual. De esta forma, la CCR coincide a su vez con la CCO correspondiente.

En este caso, el cálculo de la CCI es más complejo. Como puede haber más de una combinación de opciones correcta, la CCI equivaldría a la suma de las CCR de todos los patrones de respuesta correcta. Es decir,

$$P_i(R_i(\vec{u}_i) = 1|\theta) = \sum_{c=1}^{2^m} P_{i\vec{u}_c}(\vec{u}_c|\theta)R(\vec{u}_c) \quad (4.11)$$

4.5.4. Ítems de ordenación

En estos ítems, se muestra al alumno un conjunto de elementos, que debe ordenar según un criterio indicado. En este caso, las posibles opciones de respuesta son el espectro de posibles ordenaciones del conjunto de elementos. Asimismo, también es factible dejar la respuesta en blanco, que correspondería al caso en el que el examinando no establece ningún tipo de ordenación. Por consiguiente, sea m el conjunto de elementos del ítem, el número de posibles respuestas es igual a permutaciones de m más uno (respuesta en blanco), es decir, $m! + 1$.

Un ejemplo de este tipo de ítems, sería el siguiente: *Ordena cronológicamente (en orden ascendente) los siguientes acontecimientos históricos: a) Caída del muro de Berlín b) Segunda Guerra Mundial c) Guerra de Vietnam d) Guerra del Golfo e) Guerra de las Malvinas.* En este caso, la única ordenación correcta es la siguiente: *b, c, e, a, d.*

Las opciones de respuesta son, por tanto, las posibles ordenaciones de los m elementos, que a su vez representan el conjunto de patrones de respuesta al ítem. Esto implica que, al igual que sucede con los ítems de respuesta múltiple con opciones dependientes, las CCO van a coincidir con las CCR. Cada CCR modelará la probabilidad de que un individuo ordene los m elementos de una determinada forma, y una última CCR modelará la respuesta en blanco del alumno.

Este ítem podría considerarse un ítem de opción múltiple en el que las respuestas son las posibles ordenaciones de los elementos, y por consiguiente, podría permitirse la opción de dejar el ítem en blanco.

Por último, la CCI de este tipo de ítems, equivale a la suma de las CCR de los patrones de respuesta correspondientes a ordenaciones correctas. Como consecuencia, se permite la existencia de más de una ordenación correcta, por analogía con los ítems de respuesta múltiple con opciones dependientes. Como resultado, para calcular la CCI se aplica la ecuación 4.11.

4.5.5. Ítems de relación

En este tipo de ítems, se muestran el enunciado y dos conjuntos de elementos. El objetivo es que el alumno identifique los pares de elementos, uno de cada conjunto, que satisfacen la relación descrita en el enunciado. Por consiguiente, una opción de respuesta se corresponderá con un par de elementos de ambos conjuntos. Como consecuencia, un patrón equivaldrá a un conjunto de pares.

Formalmente, llámese A al primer conjunto de elementos del ítem, $A = \{a_1, a_2, \dots, a_{\zeta(A)}\}$, que estará compuesto por $\zeta(A)$ elementos; y B al segundo conjunto de elementos del ítem, $B = \{b_1, b_2, \dots, b_{\zeta(B)}\}$, que estará formado a su vez por $\zeta(B)$ elementos. En virtud de la cardinalidad de la relación entre los integrantes de ambos conjuntos, podrán definirse los siguientes tipos de ítems: de emparejamiento y de asociación, que se explican a continuación.

Ítems de emparejamiento

En los *ítems de emparejamiento* (*de asociación de uno a uno o de correspondencia*), cada elemento del primer conjunto puede estar relacionado con un único elemento del segundo y viceversa. Se trata de una función inyectiva, y como consecuencia, la cardinalidad de ambos conjuntos debe ser la misma, esto es, $\zeta(A) = \zeta(B)$.

Un ejemplo de un ítem de este tipo sería el siguiente: *Sea A un conjunto de eventos acaecidos en el transcurso del siglo XX, $A = \{\text{Inicio de la I Guerra Mundial, Inicio de la II Guerra Mundial, Llegada del hombre a la Luna, Inicio de la Guerra Civil Española}\}$, y B un conjunto de años del siglo XX, $B = \{1969, 1947, 1939, 1936\}$, asocia cada elemento del conjunto A con el elemento adecuado del conjunto B.* La solución correcta sería el siguiente conjunto de pares de elementos:

$$\{(\text{Inicio de la I Guerra Mundial}, 1914); (\text{Inicio de la II Guerra Mundial}, 1939); \\ (\text{Llegada del hombre a la Luna}, 1969); (\text{Inicio de la Guerra Civil Española}, 1936)\}$$

En este tipo de ítems, el conjunto de patrones de respuesta pueden verse como todas las posibles ordenaciones de uno de los dos conjuntos con respecto al otro. Es decir, si se

mantiene la ordenación de, por ejemplo, los elementos de A , las posibles soluciones serían todas las permutaciones de los elementos de B . Cada solución equivaldría a los pares de los elementos de ambos conjuntos situados en la misma posición. Por consiguiente, este tipo de ítems son muy similares a los de ordenación, y su tratamiento desde el punto de vista del modelo de respuesta, es el mismo. Por este motivo, las opciones de respuesta coinciden con las posibles respuestas, y como resultado las CCR coincidirán con las CCO. Esto se explica, porque las opciones de respuesta (los pares de elementos) no son independientes entre sí. La selección de un par de elementos condiciona el resto de pares elegidos. De esta forma, el número total de patrones de respuestas coincide con el número de emparejamientos que se pueden establecer, además de la posible respuesta en blanco. Por consiguiente, el número total de CCR (y de CCO) es igual a $m! + 1$.

Por último, por analogía con los ítems de ordenación, la CCI del ítem corresponderá con la CCR del patrón de respuesta correcto, se calcula aplicando la ecuación 4.11.

Ítems de asociación

En los *ítems de asociación*, los elementos de un conjunto pueden estar relacionados con uno o más elementos del otro. En este caso, la cardinalidad de ambos conjuntos podrá ser diferente.

Un ejemplo de un ítem de este tipo sería el siguiente: *Enlaza cada plato típico malagueño del siguiente conjunto, $A = \{gazpacho, porra, migas, ajoblanco, gazpachuelo\}$, con sus ingredientes principales, de entre los enumerados en el siguiente conjunto $B = \{agua, ajo, tomate, aceite, pan, almendras\}$* . La respuesta correcta sería el siguiente conjunto de pares:

$$\{(gazpacho, agua); (gazpacho, ajo); (gazpacho, tomate); (gazpacho, aceite); (gazpacho, pan); (porra, ajo); (porra, aceite); (porra, tomate); (porra, pan); (migas, agua); (migas, ajo); (migas, aceite); (migas, pan); (ajoblanco, agua); (ajoblanco, ajo); (ajoblanco, pan); (ajoblanco, almendras)\}$$

Este tipo de ítems son equivalentes a los de respuesta múltiple con opciones independientes. Consecuentemente, cada uno de los posibles emparejamientos tendrá asociado una CCO diferente. Formalmente, para este tipo de ítems, una opción de respuesta r_j equivaldrá a un determinado par, $r_j = (a_d, b_e)$, donde $a_d \in A$, y $b_e \in B$; siendo $1 \leq d \leq \zeta(A)$, y $1 \leq e \leq \zeta(B)$. El conjunto de todas las posibles opciones de respuesta es igual a:

$$\{(a_1, b_1); (a_1, b_2); \dots; (a_1, b_{\zeta(B)}); (a_2, b_1); (a_2, b_2); \dots; (a_2, b_{\zeta(B)}); \dots; (a_{\zeta(A)}, b_1); (a_{\zeta(A)}, b_2); \dots; (a_{\zeta(A)}, b_{\zeta(B)})\}$$

Este conjunto será de tamaño $m = \zeta(A) \times \zeta(B)$; valor que coincidirá, por tanto, con el número de CCO de un ítem de este tipo. El cálculo de la CCR correspondiente al patrón de respuesta seleccionado por el alumno vendrá definido por la ecuación 4.9; y el de la CCI por la ecuación 4.10.

Una vez estudiados los distintos tipos de ítems que ofrece este modelo, es necesario reseñar que los criterios que marcan las pautas a seguir en la construcción de un banco de ítems, recomiendan restringir el número de opciones de respuesta que deben acompañar a un determinado ítem. En general, lo recomendable, según Osterlind (1998), es que un ítem

tenga entre tres y cinco alternativas. Además, para garantizar las condiciones de la TRI, es necesario que las opciones de respuesta sean independientes entre sí (excepto en aquellos ítems en los que las CCO y CCR coinciden), y que, por tanto, seleccionar una de ellas no implique la necesidad de elegir alguna de las demás.

A partir de la clasificación anterior, es posible representar, mediante el modelo propuesto, otros tipos de ítems. Por consiguiente, los ítems que han sido descritos en esta sección pueden verse más que como tipos de ítems, como esquemas de evaluación. Esto implica que cualquier tipo nuevo que pueda evaluarse como alguno de los anteriormente descritos es, por tanto, abordable mediante este modelo de respuesta. Como se verá en el capítulo 6, el sistema que implementa el modelo de diagnóstico descrito en esta tesis, incluye otros tipos de entre aquéllos que fueron definidos en la sección 2.5.

Por último, obsérvese que cuando cualquiera de los ítems descritos en esta sección se trata como si fuera uno verdadero/falso, se obtiene una versión dicotómica de este modelo de respuesta.

4.6. Aproximación cuasipolítómica del modelo de respuesta

Muchos de los ítems anteriores no son viables desde el punto de vista computacional. La mayoría implican el cálculo de un número excesivamente grande de CCR. Por ejemplo, los ítems de respuesta múltiple dependiente requieren calibrar 2^m CCR (siendo m el número de opciones de respuesta); y los de ordenación, $m! + 1$ CCR (siendo m el número de elementos a ordenar). Aunque se dispusiera de la cantidad de información necesaria para calibrar esas curvas, los problemas no acaban ahí. Como se verá en el capítulo siguiente, algunos de los criterios de selección desarrollados para el modelo de diagnóstico, necesitan de todas las CCR del ítem para determinar durante un test si ese ítem es el más adecuado para ser presentado al alumno. Esto supone que el rendimiento computacional para aplicar el criterio de selección podría verse mermado considerablemente.

Las limitaciones anteriores entran en confrontación directa con uno de los objetivos primordiales de esta tesis, esto es, desarrollar un modelo de diagnóstico cognitivo que realmente pueda implementarse y que funcione de forma eficiente para inferir el conocimiento del alumno dentro de un STI. Por este motivo, se ha desarrollado una versión del modelo, cuyo objetivo es buscar un compromiso entre las características polítómicas y su factibilidad desde el punto de vista computacional. El resultado obtenido es la que se ha denominado *aproximación cuasipolítómica* o *parcialmente polítómica* del modelo de respuesta.

Mediante esta aproximación se reduce el número de CCR asociadas a cada ítem. El procedimiento que se sigue para determinar qué CCR se van a considerar y cuáles se descartan se basa en aplicar *bootstrapping* a los datos empleados en la calibración.

Como mencionó en la sección 2.9.2, los procedimientos de calibración de ítems utilizan la información de sesiones de tests realizados con los ítems cuyas curvas características se desea inferir. La calibración es un procedimiento estadístico que únicamente utiliza (de esa información de entrada) la puntuación obtenida por cada alumno en el test, y para el caso polítómico qué opciones de respuesta seleccionó. La idea que se intenta explotar, es que esas sesiones contienen información adicional que indica qué patrones de respuesta seleccionan los examinandos con más frecuencia. Precisamente estos datos se utiliza en la aproximación cuasipolítómica para determinar qué CCR van a ser modeladas. Así, para cada ítem sólo se

almacenarán las CCR de los patrones de respuesta más frecuentes, y todos los restantes se modelarán con una única curva. De este modo, sea m el conjunto de patrones de respuesta para un determinado ítem i , y sea b el número de patrones de respuestas más frecuentes, $b < m$, el número de CCR que tendrá asociadas el ítem i será igual a $b + 1$.

Como consecuencia, mediante la aplicación de la aproximación cuasipolítómica, se resuelve el problema de tener que calibrar CCR cuyos patrones de respuesta han sido seleccionados por un número pequeño de individuos de la muestra poblacional. Igualmente, se reduce el problema de ineficiencia computacional, puesto que al disminuirse el número de CCR, el cálculo de los criterios de selección de ítems es más factible.

Aunque la aproximación cuasipolítómica del modelo de respuesta puede aplicarse a todos los tipos de ítems descritos anteriormente, ésta es especialmente útil en aquéllos que requieren un mayor número de CCR, esto es, los de respuesta múltiple con opciones dependientes, los de ordenación y los de emparejamiento.

4.7. Calibración de las curvas características

La calibración de las curvas características es un aspecto importante a la hora de definir un modelo de respuesta, puesto que, para que éste sea realmente útil, debe disponer de un procedimiento que permita inferir los valores asociados a las curvas características. Sin un método de calibración adecuado, los modelos de respuesta son impracticables y, por lo tanto, inútiles. Para el modelo propuesto, es necesario disponer de un algoritmo que permita determinar las CCO de cada ítem.

En la mayoría de los mecanismos de calibración existentes se administra un test (no basado en la teoría de los TAI) a un conjunto de alumnos con los ítems que se desea calibrar. Esto significa que todos los examinandos realizan un test del mismo tamaño y con los mismos ítems. Éstos son evaluados siguiendo un criterio heurístico convencional como, por ejemplo, el porcentaje de ítems correctamente respondidos.

En muchas ocasiones, cuando se calibran ítems aplicando modelos paramétricos, no se suele tener en cuenta si aquéllos se pueden ajustar realmente a ese tipo de modelos. Según Stout (2001) esto es inaceptable desde el punto de vista de la estadística y la teoría de la medida moderna. Como se puso de manifiesto al comienzo de este capítulo, las curvas características no siempre pueden ser modeladas mediante aproximaciones paramétricas (Douglas y Cohen, 2001). Este problema ha hecho que muchos investigadores se inclinen por la investigación teórica y las aplicaciones de la TRI no paramétrica y, en general, por métodos de estimación de curvas características que no se restringen a funciones descritas a través de parámetros.

En la sección 2.9.2 se mostraron los métodos de calibración más comúnmente utilizados para modelos de respuesta paramétricos. Su aplicación dos principales inconvenientes:

- Requieren de una muestra poblacional bastante amplia para llevar a cabo la calibración. Para el caso de modelos polítómicos este requisito convierte al modelado de este tipo de ítems en algo prácticamente inabordable.
- Se trata de métodos iterativos que requieren un tiempo de cómputo considerable, incluso pudiendo no llegar a un resultado coherente.

El segundo inconveniente es más fácilmente evitable gracias a la potencia de cómputo que poseen los ordenadores modernos. El primero es más difícil de solucionar: los grandes requisitos que a priori imponen los TAI son unos de los principales motivos que justifican que, aún siendo tan potentes, no se utilicen tanto como cabría esperar. En general, su ámbito de aplicación queda reducido a grandes organismos educacionales, como el norteamericano *Educational Testing Service*, que con un gran número de examinandos son capaces de llevar a cabo calibraciones de forma adecuada, y consecuentemente de administrar TAI. Es por tanto fundamental, buscar técnicas alternativas que, a partir de muestras poblacionales de tamaño menor, permitan realizar calibraciones con resultados razonables.

En la sección 2.9.2 se presentó el *suavizado núcleo* como una técnica de calibración no paramétrica, que permite tratar fácilmente modelos dicotómicos y politómicos. La idea principal del suavizado núcleo es obtener una estimación no paramétrica de las curvas características tomando una media ponderada en cada punto de evaluación, en la que los pesos se determinan mediante una función núcleo (Ferrando, 2004). Douglas (1997) mostró que cuando se aplican técnicas de suavizado para la calibración de ítems, el mayor error de estimación para todos ellos converge a cero con probabilidad uno, conforme crecen el número de ítems y el de individuos utilizados en la calibración. Asimismo, para estudiar si la función 3PL era adecuada para modelar ítems, Lord (1970) necesitó 100.000 sesiones de tests, a partir de las cuales obtuvo una estimación no paramétrica. Aplicando técnicas de suavizado estadístico, el mismo análisis podría haber requerido únicamente 500 sesiones o incluso un número menor (Douglas y Cohen, 2001).

Por otro lado, en (Douglas, 1999) se muestra que para tests con un número considerable de ítems, si las curvas características son estimadas por métodos diferentes, los resultados deben ser casi idénticos. Consecuentemente, en ese caso, es posible asegurar que si existe una única estimación para cada curva característica, un método no paramétrico podrá estimarla de forma consistente. Por esta razón, según Douglas y Cohen (2001), si se llevan a cabo calibraciones utilizando aproximaciones no paramétricas y paramétricas, y existen discrepancias considerables entre ambos resultados, es posible concluir que el modelo paramétrico elegido no se ajusta correctamente a los datos.

Para el modelo de respuesta propuesto en esta tesis, se ha realizado una adaptación (Guzmán y Conejo, 2005) del método de calibración propuesto originalmente por Ramsay. Tras exhaustivos estudios empíricos, se han modificado algunas de las fases del algoritmo original ya que, de esta forma, y para este modelo, se obtenían mejores resultados que los conseguidos con la propuesta original de Ramsay, tal y como se pondrá de manifiesto en el capítulo 7.

Las fases de las que consta este algoritmo de calibración para el modelo de respuesta descrito en este capítulo, se han esquematizado en la figura 4.1, y se enumeran a continuación:

1. *Cálculo de las evaluaciones*: Para cada sesión de evaluación se calcula su calificación asociada.
2. *Conversión de las evaluaciones*: La evaluación asignada a cada sesión es sometida a un proceso de normalización.
3. *Ordenación de las sesiones*: Se ordenan las sesiones en orden creciente, según el valor asignado a la evaluación del alumno, tras haberle aplicado la transformación.
4. *Aplicación del suavizado*: Aquí se lleva a cabo la calibración propiamente dicha de cada curva característica, aplicando suavizado núcleo.

5. *Estimación de los niveles de conocimiento:* Tras calibrar las curvas características, se estima el nivel de conocimiento en el test de cada alumno de la muestra, según el modelo de respuesta.
6. *Refinamiento iterativo:* Los pasos anteriores se repiten hasta que las estimaciones, tanto de las curvas características como de los niveles de conocimiento, se estabilicen.

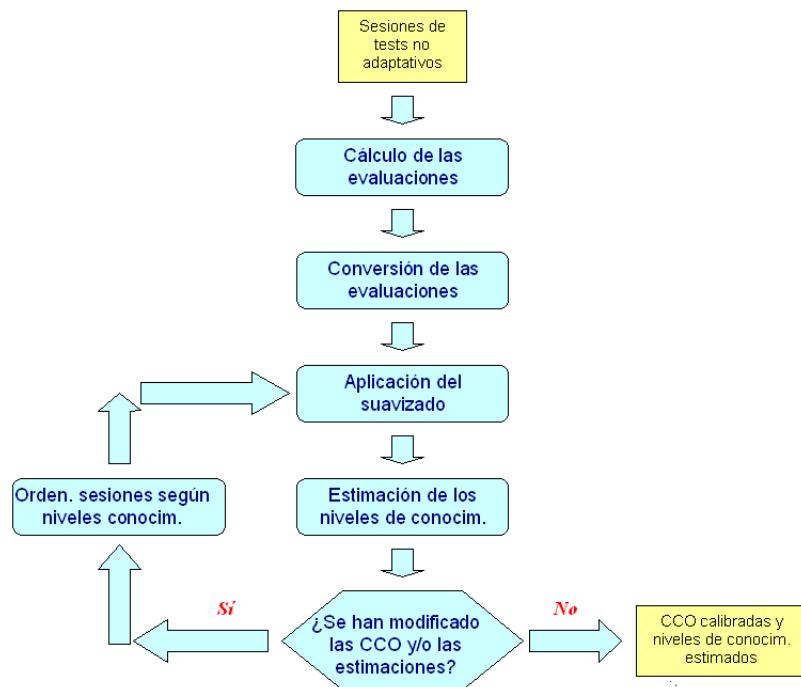


Figura 4.1: Algoritmo de calibración de las CCO.

A continuación se describirán las fases más importantes del algoritmo. Como se trata de la adaptación de una propuesta previa, en cada fase se establecerá la analogía con la aproximación original de Ramsay, y como consecuencia, se intentará poner énfasis en las mejoras introducidas.

Cálculo de las evaluaciones

Las sesiones de tests realizadas por alumnos de forma convencional, deben evaluarse siguiendo algún criterio. En general, el método de evaluación que se suele utilizar es el basado en el porcentaje de ítems correctamente respondidos. Otro criterio utilizado es el del porcentaje de ítems respondidos correctamente, pero sin contar al ítem que está siendo calibrado.

Al coexistir en este modelo diversos tipos de ítems, el método heurístico de evaluación que se ha seleccionado para esta fase del algoritmo, es el criterio por puntos, en el que cada ítem tiene asignado un número de puntos, de tal forma que la calificación de un alumno será igual a la suma de los que ha obtenido en cada ítem. Para determinar la bondad del resultado,

esta puntuación será comparada con la que se obtendría si todos los ítems hubieran sido respondidos de forma correcta. La ventaja que aporta este heurístico de evaluación frente a los anteriores, es que permite asignar valores distintos de cero a aquellos ítems respondidos de forma parcialmente correcta.

En la aplicación que se hace de este criterio de evaluación, cada ítem recibe un punto si es respondido correctamente. Para los de verdadero/falso y los de opción múltiple, cualquier otra respuesta tendrá puntuación cero. Para los de respuesta múltiple el procedimiento de puntuación difiere: cada opción de respuesta tiene asignada parte de la puntuación global del ítem. Por defecto, si el ítem tiene m opciones de respuesta, la puntuación que se asigna a cada una de ellas es igual a $1/m$. Por consiguiente, una vez que un alumno selecciona un patrón, su puntuación se calcula de la siguiente forma: por cada opción correcta se suma $1/m$. Se dice que una opción es correcta, si lo es según el enunciado y el alumno la ha seleccionado; o bien si es incorrecta y no ha sido seleccionada. La suma total será, por tanto, la puntuación asignada en ese ítem. El criterio de evaluación por puntos será explicado con más detalle en la sección 5.3.1.

Esta fase del algoritmo se corresponde con la de "clasificación" de la propuesta original de Ramsay. La única diferencia reside en el heurístico utilizado. Ramsay únicamente considera ítems de opción múltiple, por lo que los heurísticos de tipo porcentual (porcentaje de ítems acertados) o similares, resultan adecuados para su calibración. Para hacer esto con los ítems según el modelo de respuesta presentado en esta tesis, resulta más adecuado utilizar un criterio por puntos, puesto que permite puntuar parcialmente a aquellos ítems que no son completamente incorrectos.

Conversión de las evaluaciones

Una vez calculadas las calificaciones de los examinados en el test anterior, éstas son inicialmente ordenadas de forma creciente, y posteriormente transformadas. Lo más frecuente según Ramsay (2000), Douglas y Cohen (2001) es reemplazar el valor de la evaluación por el cuantil de la distribución normal estándar de la muestra poblacional. Es decir, se divide la distribución normal en $N + 1$ partes iguales, siendo N el tamaño de la muestra. A continuación, se sustituye el percentil correspondiente a la evaluación del alumno (según su posición en la ordenación de las calificaciones), por su valor en la distribución normal estándar. Por ejemplo, si la calificación obtenida corresponde al percentil 95, ésta se sustituiría por 1,645.

Esta fase del algoritmo de calibración se corresponde con la denominada "enumeración" en la propuesta original de Ramsay. La única diferencia que se ha introducido es la siguiente transformación adicional:

Una vez realizada la conversión de las calificaciones heurísticas, éstas serán números reales pertenecientes al rango $[-2, 5, 2, 5]$. Puesto que en el modelo de respuesta discreto propuesto, los niveles de conocimiento se miden utilizando números naturales entre cero y el número de niveles de conocimiento menos uno, es necesario realizar una transformación (discretización) que exprese esos niveles reales en valores naturales (discretos) dentro del rango empleado.

Aplicación del suavizado

Esta fase corresponde al "suavizado" del algoritmo original de Ramsay. Como consecuencia, en ella es donde se realiza realmente la calibración, y por lo tanto, por cada ítem se calcularán sus CCO.

Sea N el número de sesiones de tests convencionales realizadas, y k una variable utilizada para representar un nivel de conocimiento, la CCO de la opción j de un ítem i se calcula de la siguiente forma:

$$\forall k, \quad k \in \{0, 1, 2, \dots, K - 2, K - 1\}, \quad P_{ij}(S(r_j) = 1 | \theta = k) = \sum_{a=1}^N w_{ak} u_{ija} \quad (4.12)$$

donde $u_{ija} = 1$ indica que el alumno a-ésimo seleccionó la opción j del ítem i . En caso contrario, $u_{ija} = 0$. Además, cada peso w_{ak} se calcula de la siguiente forma:

$$w_{ak} = \frac{\kappa(\frac{\theta_a - \theta_k}{h})}{\sum_{b=1}^N \kappa(\frac{\theta_b - \theta_k}{h})} \quad (4.13)$$

siendo θ_a la calificación del alumno a-ésimo calculada en el paso anterior, θ_k el nivel de conocimiento para el cual se está calculando la probabilidad correspondiente según la CCO, κ la función núcleo, y h el parámetro de suavizado. La elección del valor más adecuado para el parámetro de suavizado es un tema muy estudiado y sobre el que existen un gran número de heurísticos. La idea que subyace en la elección del valor de h es minimizar el error cuadrático medio (ECM) de la estimación (Douglas y Cohen, 2001). Cualquier proceso de suavizado supone una búsqueda de equilibrio entre el error de la estimación y la varianza de la muestra utilizada en el proceso de calibración, siendo el parámetro de suavizado el encargado de controlar este equilibrio (Härdle, 1992). Si h tiene un valor pequeño, el error de estimación será pequeño, ya que sólo aquellas observaciones que sean muy cercanas a θ serán ponderadas para obtener el valor de θ . Por el contrario, cuando se dispone de una muestra poblacional de menor tamaño (la varianza de la distribución de la muestra es mayor), deben tenerse en cuenta un mayor número de observaciones para estimar θ , incrementándose en este caso el valor de h , y obteniendo como resultado estimaciones más sesgadas, es decir, con un error de precisión mayor. Como se vio en la sección 2.9.2, para calcular el valor de h , Ramsay (1991) utiliza el siguiente heurístico: $h = 1,1N^{-1/5}$. En el modelo propuesto en este capítulo, tras realizar un estudio empírico exhaustivo, se ha llegado a la conclusión de que el valor más adecuado del parámetro de suavizado oscila entre 0,75 y 0,85, no mostrando este valor sensibilidades notables al cambio en el tamaño de la muestra de sujetos utilizada en la calibración.

Estimación de los niveles de conocimiento

Una vez aplicado el suavizado, se obtienen como resultado las CCO de todos los ítems del test. Ahora es necesario llevar a cabo la estimación del conocimiento de los alumnos de la muestra, pero no de forma heurística, tal y como se hizo al principio del algoritmo, sino aplicando el propio modelo de respuesta.

Así, por cada alumno se calcula su nivel de conocimiento real aplicando el método de estimación de máxima verosimilitud. Anteriormente, en la fórmula 2.28, se definió cómo se calcula la máxima verosimilitud en los modelos dicotómicos. Para su aplicación a modelos politómicos, basta con sustituir la curva característica de cada ítem, por la CCR del patrón correspondiente a la respuesta elegida. Sea por tanto, \vec{u}_n el patrón de respuestas del alumno en los n ítems del test de calibración, $\vec{u}_n = \{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_n\}$, la máxima verosimilitud se calcula aplicando la siguiente ecuación:

$$P(\theta|\vec{\mathbf{u}_n}) = \prod_{i=1}^n P_{i\vec{\mathbf{u}_i}}(\vec{u_i}|\theta) \quad (4.14)$$

Tras la aplicación de la máxima verosimilitud se obtiene una distribución del conocimiento. A partir de ella, el nivel de conocimiento estimado del alumno equivale al máximo de la distribución resultante.

Refinamiento iterativo

Aunque el procedimiento de calibración podría darse por terminado en este momento, es posible mejorar los resultados obtenidos aplicando un proceso de refinamiento iterativo. Una vez realizada la primera estimación de las CCO, y utilizando también las nuevas estimaciones del nivel de conocimiento de los alumnos, se vuelve a aplicar el proceso de calibración de las CCO. El procedimiento se va repitiendo hasta la estimación se estabilice, es decir, hasta que los resultados obtenidos en la calibración y en la posterior inferencia del nivel de conocimiento de los alumnos no varíen sustancialmente.

Inicialmente, la propuesta de Ramsay sugiere la aplicación adicional de este paso. Así, el objetivo era, partiendo de estos nuevos resultados de evaluación de los examinandos, volver a aplicar el algoritmo de calibración desde el principio. Tras diversos estudios empíricos, y tal y como se mostrará en el capítulo 7, se ha concluido que, al menos para este modelo, si el refinamiento se aplica a partir de tercer paso se obtienen estimaciones notablemente mejores. Es decir, en la propuesta inicial, tras el primer suavizado y la posterior estimación de los niveles de conocimiento, las nuevas calificaciones obtenidas eran convertidas según el procedimiento indicado en el paso dos. Tras el estudio empírico que se ha realizado, se ha descubierto que se consiguen mejores resultados cuando se mantienen las estimaciones del conocimiento del alumno, en vez de sustituirlas por calificaciones heurísticas y así, tras ordenarlas, se vuelve a aplicar el suavizado directamente, tal y como se muestran en la figura 4.1.

Finalmente, para determinar qué función núcleo de las mostradas en la sección 2.9.2 es la más adecuada para este algoritmo de calibración, se realizó un estudio empírico en el que a partir de diversos conjunto de sesiones de tests simuladas, se calibraron sus ítems. Se realizaron tres calibraciones diferentes, cada una de ellas utilizando una de las funciones núcleo, cuyos resultados mostraron que la función gaussiana (ecuación 2.53) es la más adecuada. Ésta, a diferencia de las anteriores, daba lugar a estimaciones de las CCO y de los niveles de conocimiento de los examinandos más precisas que el resto.

4.8. Discusión y conclusiones

En este capítulo se ha presentado un modelo de respuesta basado en la TRI que es *discreto*, por lo que las curvas características y las distribuciones del conocimiento estimado del alumno son vectores. Esto evita tener que utilizar un método (iterativo) de aproximación numérica para calcular su nivel, a partir de su distribución de conocimiento.

Este modelo es, además *polítómico*, con lo que cada patrón de respuesta tiene asociada una curva característica propia, la CCR. Asimismo, se ha introducido un nuevo tipo de curvas características, las CCO, que están asociadas a cada opción de respuesta. La combinación de una o más opciones forman un patrón de respuesta a un ítem; por lo tanto, a

partir de una o más CCO, se puede inferir la CCR correspondiente. La principal ventaja del uso de modelos politómicos es que requieren, en cada test, un número menor de ítems que los dicotómicos. Esto es debido a que este tipo de ítems pone el énfasis en discernir entre las distintas respuestas que puede seleccionar el alumno, a diferencia de los modelos dicotómicos, donde la respuesta se representa tan sólo mediante un valor binario (correcta o incorrecta).

Se trata también de un modelo *heterogéneo* aplicable a tests en los que pueden coexistir diversos tipos de ítems. En general, los modelos de respuesta suelen definirse para un tipo de ítems concreto o, como mucho, para ítems diferentes en cuanto a su formato, pero no en su tratamiento desde el punto de vista de la evaluación. Son pocos los tests que mezclan diversos tipos de ítems. Uno de los principales motivos, tal y como indica Thissen (1993), es la dificultad que supone evaluar tests en los que se combinan ítems con diversos formatos. Por ejemplo, aunque en los modelos dicotómicos pueden coexistir ítems verdadero/falso y de opción múltiple, desde el punto de vista de la evaluación, esta diferenciación no es relevante, puesto que ambos sólo podrán considerarse como correctos o incorrectos, sin hacer la menor distinción en términos de la respuesta seleccionada por el alumno. Este modelo, por el contrario, define diversos tipos diferentes de ítems. Su inclusión no sólo se traduce en mejoras en características de los tests, tales como la precisión de las estimaciones o el número de ítems requeridos, sino que a su vez no suponen ninguna merma en cuanto a la uniformidad de su tratamiento desde el punto de vista matemático. Esto es, una vez que el alumno responde a un ítem, para todos los tipos se sigue el mismo procedimiento: se calcula la CCR correspondiente y ésta se utiliza para inferir el conocimiento del alumno o para determinar el siguiente ítem del test.

Por último, el modelo de respuesta es *no paramétrico*, facilitando no sólo el procedimiento de calibración de los ítems, sino también reflejando la realidad del comportamiento de los alumnos frente a éstos, en vez de seguir funciones que determinen este comportamiento. Quizás una de las justificaciones más convincentes del uso de aproximaciones no paramétricas en modelos de respuesta politómicos, es la aportada por Abrahamowicz y Ramsay (1992). Según ellos, los modelos paramétricos de la TRI no ofrecen la suficiente flexibilidad para representar la variedad de relaciones presentes en los datos utilizados para calibrar los ítems según un modelo de respuesta politómico.

En cuanto al procedimiento de calibración de los ítems, se fundamenta en la aplicación de suavizado núcleo estadístico; y aunque se basa en una propuesta anterior (Ramsay, 1991), ésta ha sido modificada, para ajustarla a las características de modelo heterogéneo y discreto, obteniéndose mejores resultados que con la propuesta original, tal y como se pondrá de manifiesto en el capítulo 7.

Este modelo de respuesta basado en la TRI, es el primer paso para definir un modelo de diagnóstico cognitivo basado en TAI para la evaluación multiconceptual de los elementos de un currículo, el cual se presenta en el capítulo siguiente.

Capítulo 5

Un modelo basado en TAI para el diagnóstico en STI

*La verdadera ciencia enseña,
por encima de todo,
a dudar y a ser ignorante.*
Miguel de Unamuno

Algunos investigadores (García et al., 1998) han puesto de manifiesto que en el uso de TAI no se aprovecha al máximo la característica más significativa de este tipo de tests, esto es, la adaptación. La verdadera aportación de los TAI no está únicamente en la capacidad de adaptarse a la situación particular de cada examinando, sino en su virtualidad para conocer qué progresos realiza el alumno en su descripción, comprensión, análisis, valoración, etc. de un determinada materia. La posibilidad de que un TAI permita obtener información sobre los progresos en el aprendizaje representa un indicador clave de su verdadera naturaleza adaptativa que, incluso, va más allá de su demostrada valía psicométrica.

Según Chipman et al. (1995), en los TAI se ha llegado a un estado un tanto peculiar, tanto desde el punto de vista teórico como desde el práctico. Se ha elaborado y refinado un sofisticado aparato matemático, mediante el cual es posible seleccionar los ítems más convenientes, ensamblarlos en tests apropiados, y convertir los resultados obtenidos por un alumno en el test en escalas de medición adecuadas. Por otro lado, estos tests parecen haber sido diseñados para ordenar y comparar individuos entre sí, para calificarlos y para predecir cuáles de ellos desempeñarán mejor una tarea en el futuro. Por el contrario, no suministran información útil para el diagnóstico, concretamente sobre el contenido específico que el examinando debería estudiar (o ser enseñado), con el objetivo de mejorar sus resultados.

Los *modelos de evaluación cognitiva* (en inglés, *Cognitive Assessment Models*), en general, persiguen un objetivo más ambicioso que una nueva clasificación de forma lineal de los alumnos, tal y como suelen hacer los modelos de respuesta basados en la TRI. Los *de evaluación cognitiva* proporcionan una lista de habilidades u otros atributos cognitivos que el alumno puede poseer, en función de las evidencias proporcionadas por las tareas que realiza (Junker y Sijtsma, 2001a). En resumen, son modelos que tienen en cuenta y miden los procesos cognitivos y las estrategias seguidas para responder a ítems dicotómicos o politómicos (Junker y Sijtsma, 2001b).

A raíz de la problemática anterior, se va a proceder a presentar la contribución principal de esta tesis, tomando como base la teoría de los TAI. Se propone un modelo de evaluación mediante tests adaptativos, que aplica la teoría de los TAI al diagnóstico del alumno dentro de los STI, y utiliza el modelo de respuesta presentado en el capítulo anterior. El objetivo principal, es dotar a los STI de un sistema para la inferencia del estado del conocimiento del estudiante, con un trasfondo teórico que garantice que los resultados obtenidos en la evaluación sean certeros. El modelo propuesto podrá ser empleado para inicializar y actualizar el modelo del alumno de un STI lo que garantiza, al haber sido inferido con técnicas bien fundamentadas, que la información que contiene es correcta. Esto puede contribuir, en gran medida, a que el proceso de adaptación, llevado a cabo por el planificador de instrucción del STI, sea más preciso.

Por otra parte, a lo largo de este documento se han puesto de manifiesto las dificultades de la aplicación de los TAI al ámbito de los STI, puesto que sólo son capaces de medir un único rasgo latente cada vez. Además, cuando un test evalúa múltiples conceptos de forma simultánea, no se dispone de ningún mecanismo que garantice que la selección de los ítems esté balanceada. El modelo presentado intenta paliar estos problemas, de forma que, en una misma sesión (o lo que es lo mismo, que en un mismo test), se puedan evaluar múltiples conceptos de forma simultánea, sin perder el rigor teórico. Como se ha descrito en capítulos anteriores, la gran mayoría de las propuestas basadas en TAI con balanceo en contenido, introducen pesos determinados de forma heurística por los profesores, perdiendo como consecuencia, el rigor teórico.

Según Chipman et al. (1995), los investigadores que han hecho propuestas para abordar el problema del diagnóstico cognitivo desde una perspectiva psicométrica (Tatsuoka, 1985; Embretson, 1987; DiBello et al., 1995; Samejima, 1995; Junker y Sijtsma, 2001a) tienden a hacer un tratamiento abstracto y matemático. En el modelo que se presenta en este capítulo se hará una descripción formal, pero siempre con el objetivo de que el resultado se traduzca en un modelo implementable.

Este capítulo se estructura en dos partes fundamentales. En la primera, se presentará la arquitectura del modelo, donde se describirán sus componentes de forma detallada. En la segunda parte, se detalla el funcionamiento del modelo, es decir cómo se aplica para el diagnóstico del conocimiento del alumno.

5.1. La arquitectura del modelo

El modelo combina los elementos principales de un sistema para la generación de TAI, con aquellos componentes que debe poseer el módulo de diagnóstico de un STI. En la figura 5.1 se ha representado su arquitectura, en la que se pueden distinguir, a grandes rasgos, los siguientes componentes:

- *El módulo experto:* Contiene el conocimiento aportado por el experto (en este caso, el profesor). Se compone a su vez, tres partes: (a) Un *modelo (o mapa) conceptual*, que es una representación del dominio que incluye conceptos y las relaciones entre éstos; (b) un *banco de ítems*, que contiene los ítems para llevar a cabo el diagnóstico; y por último, (c) un conjunto de *especificaciones de tests*, que son guías de evaluación definidas por los profesores, en las que se expresan los conceptos involucrados y los parámetros que caracterizarán esas sesiones de evaluación.

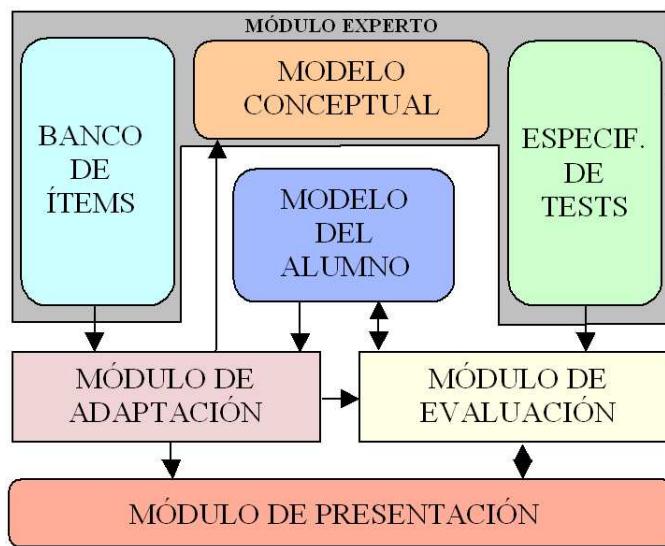


Figura 5.1: Arquitectura del modelo de diagnóstico cognitivo.

- *El modelo del alumno:* Se encarga de almacenar toda la información disponible sobre el alumno y, por tanto, su representación dentro del modelo de diagnóstico.
 - *El módulo de adaptación:* Su misión de decidir qué ítem debe mostrarse al alumno, teniendo en cuenta: (1) Información de su modelo, considerando su distribución de conocimiento estimado en los conceptos evaluados y los ítems que le han sido previamente administrados. (2) Parámetros con los que el profesor ha configurado el test: criterio de selección de ítems, conceptos evaluados, etc. (3) Conjunto de ítems que forman parte de los bancos de ítems de los conceptos evaluados en el test.
- Igualmente, este módulo es el encargado de determinar si el test debe finalizar. Para ello, antes de llevar a cabo la elección del siguiente ítem, se comprueba si se satisface el criterio de finalización. En ese caso, no se continuarán mostrando nuevos ítems, y el proceso de diagnóstico habrá concluido.
- *El módulo de evaluación:* Lleva a cabo la función de inferir el nuevo conocimiento del alumno una vez que responde a un ítem, y de actualizar su modelo en consecuencia. Asimismo, antes de tener ningún tipo de evidencia sobre lo que sabe el examinando sobre los conceptos evaluados en el test, este módulo inicializa su modelo. Posteriormente, conforme vaya respondiendo a los diferentes ítems, y en función de cada respuesta, se estimarán las nuevas distribuciones del conocimiento del alumno en los conceptos correspondientes.
 - *El módulo de presentación:* Es la interfaz de interacción con el usuario. Una vez que se ha determinado qué ítem debe ser mostrado, éste se presenta al examinando a través de este módulo. La respuesta que el alumno suministra al ítem es capturada a través de este módulo, y enviada para su evaluación. En función de la especificación del test, si procede, presenta al alumno la corrección del ítem.

5.2. Descripción del modelo

Una vez dada una descripción general de la arquitectura del modelo, a continuación se detallarán sus partes más relevantes.

5.2.1. El módulo experto

El módulo experto en los STI almacena una representación de parte del conocimiento del dominio que tiene el profesor (que es el experto en este ámbito). Este conocimiento puede expresarse mediante una red de conceptos que permita establecer una secuencia en la que los alumnos deben estudiar esos conceptos. Para poder llevar a cabo diagnósticos, es necesario disponer de los instrumentos necesarios. En este caso, estas herramientas son los ítems y las especificaciones de los tests. Como consecuencia, formalmente, el módulo experto puede verse como una tripleta (Ω, Φ, Π) compuesta por tres grupos: el conjunto de conceptos Ω , cuyos elementos están relacionados entre sí formando el modelo conceptual; el conjunto de ítems Φ ; y el conjunto de especificaciones de tests Π . En la figura 5.2 se muestra una representación gráfica de una posible estructuración del dominio, y su relación con los ítems y tests. Cada uno de estos elementos se detallará en las siguientes secciones.

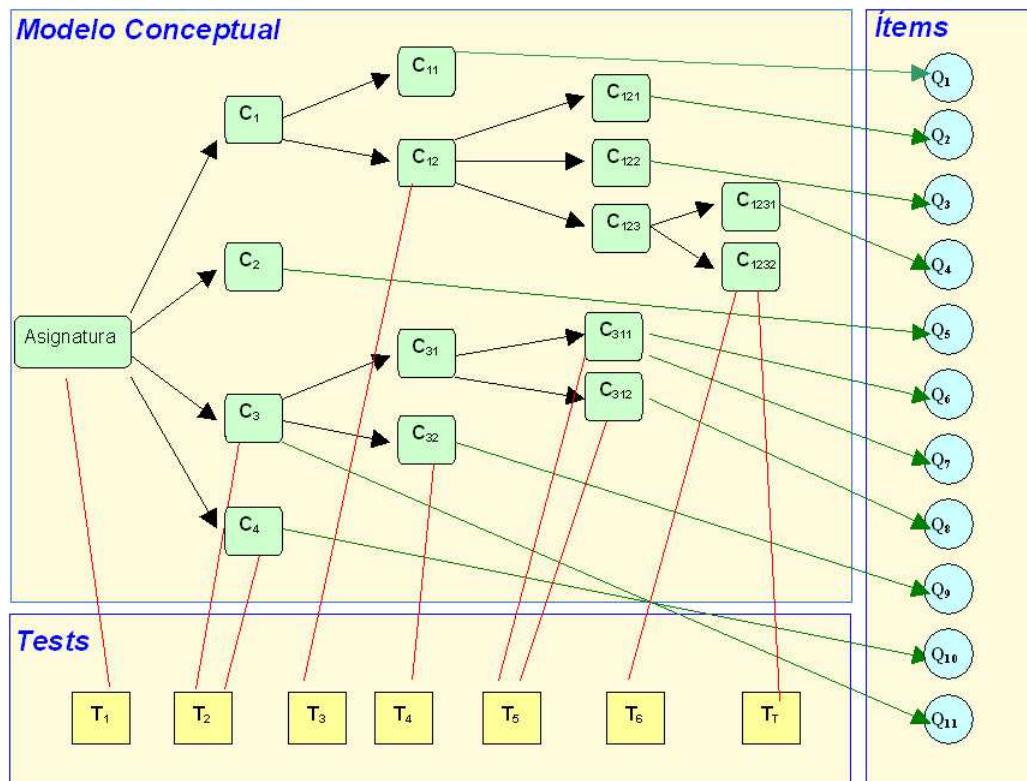


Figura 5.2: Relación entre los elementos del modelo experto.

El modelo conceptual

En los criterios tradicionales de enseñanza, para contribuir a una mejor comprensión, las materias (cursos o asignaturas) suelen estructurarse en partes, divididas a su vez en subpartes, y así sucesivamente. De esta forma, se obtienen jerarquías con granularidad variable, denominadas *currículos*. La *granularidad* de un dominio hace referencia al nivel de detalle o perspectiva desde la que los conceptos pueden ser vistos (Greer y McCalla, 1994). En el ámbito de los STI, suelen representarse mediante las denominadas *redes semánticas* (véase la sección 3.1.3), que son grafos acíclicos donde los nodos son las partes en las que se dividen las materias, y los arcos representan relaciones entre éstas. En la literatura de los STI, existen multitud de propuestas (Schank y Cleary, 1994; Rodríguez Artacho, 2000) en los que esas partes reciben nombres diferentes según su nivel dentro de la jerarquía: temas, conceptos, entidades, capítulos, secciones, definiciones, etc.

En esta propuesta, los nodos de la jerarquía recibirán el nombre genérico de *conceptos*, y la forma de distinguirlos será su profundidad dentro del árbol curricular. Se asumirá la definición proporcionada por Reye (2002), según la cual los conceptos son elementos del currículo que representan piezas de conocimiento o habilidades cognitivas que el alumno podría adquirir; es decir, se corresponde a la noción intuitiva de tema. Desde el punto de vista del diagnóstico del alumno, los conceptos serán aquellos elementos susceptibles de ser evaluados. Nótese que los nodos finales (nodos hoja) del modelo corresponden a conceptos únicos o a un conjunto de ellos indiscernibles de cara a la evaluación. Por analogía, el nodo raíz, al que se denominará *asignatura*, se considera como un concepto más, que a su vez se puede estructurar en otros, y por tanto, es posible obtener una valoración del conocimiento del alumno a nivel de la asignatura global.

En cuanto a las relaciones, se asume que los conceptos de un nivel de la jerarquía están relacionados con los niveles inmediatamente anterior y posterior mediante relaciones de agregación ("parte-de"). Es decir, se considerará que el conocimiento de un conjunto de nodos hijos, forma parte del conocimiento del nodo padre. Genéricamente, se dirá que entre los conceptos existe una *relación de inclusión*. Formalmente, esta relación se puede expresar como sigue:

Sean Ω el conjunto de conceptos de la asignatura, y Π el conjunto de ítems de la misma, se definen los siguientes tipos de relaciones entre conceptos:

Definición 5.1 (Relación de inclusión directa entre dos conceptos). Sean dos conceptos C_i y C_j , pertenecientes al conjunto de conceptos de la asignatura ($C_i, C_j \in \Omega$), se dice que existe una relación de inclusión directa entre ellos, si a su vez existe una relación de agregación entre ambos. Esto puede expresar formalmente como sigue:

$$\forall C_i, C_j, \quad C_i, C_j \in \Omega, \quad C_i \in \wp(C_j) \quad \text{si} \quad C_i \text{ está relacionado mediante} \\ \text{una relación de agregación (es parte de) con } C_j \quad (5.1)$$

Se dirá por tanto, que C_j incluye directamente a C_i . Gráficamente, existe una relación de inclusión directa entre los conceptos C_i y C_j de la red semántica, si en el grafo representativo existe un arco que une a ambos entre sí, y que va desde C_j hasta C_i . Para el modelo conceptual de la figura 5.2, entre los conceptos C_1 y C_{12} existe una *relación de inclusión directa*, puesto que $C_{12} \in \wp(C_1)$.

Definición 5.2 (Relación de inclusión indirecta entre dos conceptos). Sean los conceptos C_i y C_j , se dice que existe una relación de inclusión indirecta entre ellos, si a su vez

existe una relación indirecta de agregación entre ambos. Esto puede expresarse formalmente como indica la siguiente ecuación:

$$\begin{aligned} \forall C_i, C_j, \quad C_i, C_j \in \Omega, \quad C_i \in \wp^n(C_j) & \text{ si } \\ \exists C_k, \quad C_k \in \Omega, \quad C_k \in \wp^{n-1}(C_j) \wedge C_i \in \wp(C_k) \wedge \quad n > 1 & \end{aligned} \quad (5.2)$$

Se dirá por tanto, que C_j incluye indirectamente a C_i . Gráficamente, dos conceptos de un modelo conceptual tendrán una relación de inclusión indirecta, cuando exista un camino en el grafo representativo del modelo conceptual, que une a los nodos, con al menos un concepto intermedio. La variable n se denominará *orden de la relación indirecta*, y su valor deberá ser un número natural mayor que uno. Así, para el grafo de la figura 5.2, entre los conceptos C_1 y C_{121} existen una relación indirecta de orden 2, ya que $C_{121} \in \wp^2(C_1)$. Asimismo, entre los conceptos C_1 y C_{1231} existen una relación indirecta de orden 3, ya que $C_{1231} \in \wp^3(C_1)$.

Definición 5.3 (Relación de inclusión entre dos conceptos). Esta función es una generalización de las dos anteriores, y se define de la siguiente forma:

$$\forall C_i, C_j, \quad C_i, C_j \in \Omega, \quad C_i \in \wp^n(C_j) \iff \begin{cases} C_i \in \wp(C_j) & \text{si } n = 1 \\ \exists C_k, \quad C_k \in \Omega, C_k \in \wp^{n-1}(C_j) \wedge C_i \in \wp(C_k) & \text{si } n > 1 \end{cases} \quad (5.3)$$

Por lo tanto, dos conceptos estarán relacionados entre sí, si entre ellos existe una relación de inclusión, bien sea directa o indirecta. Nótese que cuando el orden de la relación es igual a uno, la relación de inclusión entre conceptos es directa.

Propiedades: La relación de inclusión entre conceptos tiene las siguientes propiedades:

1. *Asimétrica:* Esto es, si $C_i \in \wp^n(C_j) \Rightarrow C_j \notin \wp^n(C_i)$. Por lo tanto, se trata de una relación direccional entre conceptos. Es decir, la relación tiene una única dirección.
2. *No reflexiva:* $C_i \notin \wp^n(C_i)$.
3. *Transitiva:* Si $C_i \in \wp^n(C_j) \wedge C_j \in \wp^m(C_k) \Rightarrow C_i \in \wp^{n+m}(C_k)$.

Si la relación de inclusión no es directa, es decir, si no existe un arco directo entre los dos conceptos, aunque sí existe un camino que va de C_i a C_j , por definición se dirá que existe una relación de inclusión indirecta entre los conceptos C_i y C_j . Sea $n - 1$ el número de conceptos intermedios entre C_i y C_j , esta relación de inclusión indirecta se expresa de la siguiente forma: $C_j \in \wp^n(C_i)$.

En general, desde el punto de vista de la evaluación, se dirá que si un alumno sabe el concepto C_j tendrá, en virtud de las relaciones de inclusión, una cierta noción del concepto C_i . En este modelo no se considera la posibilidad de tratar relaciones entre conceptos que estén a distinto nivel en la jerarquía curricular. Por ejemplo, no se considera la existencia de relaciones de prerrequisito; es decir, se asume que el conocimiento del examinando en un concepto es completamente independiente de su conocimiento sobre el resto de conceptos del mismo nivel. En resumen, el modelo propuesto lleva a cabo únicamente medidas del conocimiento de forma directa a partir de evidencias.

El banco de ítems

Los ítems utilizados en TAI se almacenan en los denominados banco de ítems. Aunque el concepto de banco de ítem ya se introdujo en el capítulo 2, a continuación se incluye la definición realizada por Barbero (1999):

La concepción de lo que es un banco de ítems ha ido cambiando a lo largo de los años, aunque la idea subyacente ha sido siempre la misma: un conjunto más o menos numeroso de ítems, que miden el mismo rasgo o habilidad y que se almacenan de tal manera que, llegado el momento, se pueda elegir de entre todos ellos los que mejor se adapten a las necesidades de uso.

En el modelo propuesto, los ítems se utilizan como instrumentos para el diagnóstico del conocimiento del alumno en uno o más conceptos. Los ítems son, por tanto, entidades suministradoras de evidencias de cuánto sabe; ya que a través de ellos, el modelo interactúa con el alumno. En este modelo, cada concepto tendrá asociado un banco de ítems.

A continuación se va a proceder a formalizar las relaciones entre ítems y conceptos. Como primer paso, se definirá la relación de asociación de un ítem a un concepto de la siguiente forma:

Definición 5.4 (Asociación de un ítem a un concepto). Sean un ítem Q_i , perteneciente al conjunto de ítems de una asignatura ($Q_i \in \Pi$), y un concepto C_j , perteneciente al conjunto de conceptos ($C_j \in \Omega$), se dice que Q_i está asociado a C_j , si se requiere conocer C_j para resolver el ítem Q_i . Es decir, la respuesta seleccionada por un alumno en ese ítem permite realizar inferencias sobre su nivel de conocimiento en ese concepto. Para representar esta relación se define la función $A : \Pi \times \Omega \rightarrow \{0, 1\}$ de la siguiente forma:

$$A(Q_i, C_j) = \begin{cases} 1 & \text{si } Q_i \text{ está asociado a } C_j \\ 0 & \text{en otro caso} \end{cases} \quad (5.4)$$

En la figura 5.2, se ha representado esta relación de asociación mediante una línea que une al ítem con el concepto. Así, el ítem Q_1 está asociado al concepto C_{11} , es decir, $A(Q_1, C_{11}) = 1$. Durante la construcción del módulo experto, el profesor habrá añadido el ítem Q_1 al banco de ítems de C_{11} .

Nótese que el modelo presentado restringe la relación entre ítems y conceptos a una relación unidimensional. Aunque como se puso de manifiesto en la sección 2.8.1 existen modelos de respuesta basados en la TRI en los que se contempla esta posibilidad, actualmente, en el modelo propuesto en esta tesis se asume que los ítems son siempre unidimensionales. Ciertamente, podría darse el caso de que un ítem estuviera asociado a más de un concepto del mismo nivel del currículo; es decir, para resolver el ítem el examinando debería poseer conocimientos sobre más de un concepto. Así, por ejemplo, supóngase que para resolver el ítem Q_4 del currículo de la figura 5.2 fuera necesario saber los conceptos C_{1231} y C_{1232} . Siguiendo este modelo, lo que debería hacerse es asociar ese ítem al concepto padre de ambos, esto es, a C_{123} . Al ser C_{123} una agregación de C_{1231} y C_{1232} , si el ítem se asocia al concepto padre C_{123} , se está expresando (en el módulo experto) el hecho de que el ítem proporciona evidencias sobre el conocimiento del alumno en ambos conceptos. Ciertamente, estas evidencias no permiten (para este modelo unidimensional) inferir el conocimiento del alumno directamente en C_{1231} y C_{1232} , sino en su agregación más inmediata.

Definición 5.5 (Evaluación directa de un ítem sobre un concepto). Sean el ítem Q_i y el concepto C_j , se define la función de evaluación directa de un ítem sobre un concepto, $E_D : \Pi \times \Omega \rightarrow \{0, 1\}$, de la siguiente forma:

$$E_D(Q_i, C_j) = \begin{cases} 1 & \text{si } A(Q_i, C_j) = 1 \\ 0 & \text{en otro caso} \end{cases} \quad (5.5)$$

Un ítem evaluará directamente a un concepto, cuando esté asociado a él, $A(Q_i, C_j) = 1$. Por ejemplo, como se puede apreciar en la figura 5.2, Q_6 y Q_7 evalúan directamente a C_{311} .

El concepto que un ítem evalúa de forma directa, será aquél situado a mayor profundidad en el árbol curricular, que es capaz de proporcionar una evidencia sobre el conocimiento que el alumno tiene de ese concepto.

Axioma 5.1. Todos los ítems del banco del módulo experto construido para una determinada asignatura, deben evaluar directamente a un único concepto.

$$\forall Q_i, \quad Q_i \in \Pi \implies \exists! C_j, \quad C_j \in \Omega, \quad E_D(Q_i, C_j) = 1 \quad (5.6)$$

Axioma 5.2. Considérese que el ítem Q_i posee un conjunto de opciones de respuesta $M_i = \{r_1, r_2, \dots, r_m\}$. Si Q_i evalúa directamente al concepto C_j , esta relación se substancia a través de tantas CCO como opciones de respuesta tenga el ítem. Cada CCO representa la probabilidad de que el examinando seleccione esa opción de respuesta, dado su nivel de conocimiento en C_j .

$$\forall Q_i, C_j, \quad Q_i \in \Pi, \quad C_j \in \Omega, \quad E_D(Q_i, C_j) = 1 \Leftrightarrow \forall r_m, \quad r_m \in M_i, \quad \exists P_{im}(S(r_m) = 1 | \theta_j), \quad \sum_{k=0}^{K-1} P_{im}(S(r_m) = 1 | \theta_j = k) \neq 0 \quad (5.7)$$

Es decir, un ítem Q_i evaluará directamente a un concepto C_j , cuando exista una curva característica (con al menos un valor distinto de cero) que modele la probabilidad de que un alumno responda correctamente al ítem, dado su nivel de conocimiento θ_j en C_j .

Este axioma intenta enfatizar que, cuando un ítem evalúa un concepto directamente, siempre se podrá inferir una función probabilística que represente la probabilidad de que un examinando seleccione una determinada opción del ítem, dado su nivel de conocimiento en ese concepto.

En virtud de la relación que se establece entre los conceptos que componen el modelo conceptual, un ítem podrá evaluar simultáneamente y de forma indirectamente a diversos conceptos (Guzmán y Conejo, 2002a). La relación de evaluación indirecta se define de la siguiente forma:

Definición 5.6 (Evaluación indirecta de un ítem sobre un concepto). Sean el ítem Q_i y el concepto C_j , se define la función de evaluación indirecta de un ítem sobre un concepto, $E_I : \Pi \times \Omega \rightarrow \{0, 1\}$, de la siguiente forma:

$$E_I(Q_i, C_j) = \begin{cases} 1 & \text{si } A(Q_i, C_j) = 0 \wedge \exists C_l, \quad C_l \in \Omega, \quad A(Q_i, C_l) = 1 \wedge C_l \in \wp^n(C_j) \\ 0 & \text{en otro caso} \end{cases} \quad (5.8)$$

Es decir, Q_i evaluará indirectamente a C_j , cuando exista otro concepto C_l evaluado directamente por Q_i , y cuando además, entre ambos exista una relación de inclusión. Esta última

se establece de tal forma que C_l deberá estar incluido en C_j ($C_l \in \wp^n(C_j)$). Obsérvese que, a diferencia de la evaluación directa, el ítem no está asociado al concepto, es decir, no pertenece de forma directa a su banco de ítems.

Axioma 5.3. Considérese de nuevo que el ítem Q_i posee el conjunto de opciones de respuesta $M_i = \{r_1, r_2, \dots, r_m\}$. Si Q_i evalúa indirectamente a C_j , esta relación se substancia a través de tantas CCO como opciones de respuesta tiene el ítem. Cada CCO representa la probabilidad de que el alumno seleccione esa opción de respuesta, dado su nivel de conocimiento en C_j .

$$\begin{aligned} \forall Q_i, C_j, \quad Q_i \in \Pi, \quad C_j \in \Omega, \quad E_I(Q_i, C_j) = 1 \Leftrightarrow \\ \forall r_m, \quad r_m \in M_i, \quad \exists P_{im}(S(r_m) = 1 | \theta_j), \quad \sum_{k=0}^{K-1} P_{im}(S(r_m) = 1 | \theta_j = k) \neq 0 \end{aligned} \quad (5.9)$$

Es decir, Q_i evaluará directamente a C_j , cuando exista una curva característica (con al menos un valor distinto de cero), que modele la probabilidad de que un individuo responda correctamente al ítem, dado su nivel de conocimiento θ_j en C_j .

La existencia de un conjunto de CCO se traduce en que es posible inferir (o cuantificar) la relación existente entre el conocimiento del alumno sobre el concepto y la respuesta que selecciona en el ítem.

Definición 5.7 (Evaluación de un ítem sobre un concepto). Generalizando, se define la evaluación de Q_i sobre C_j , como una función $E : \Pi \times \Omega \rightarrow \{0, 1\}$, que viene descrita de la siguiente forma:

$$E(Q_i, C_j) = E_D(Q_i, C_j) + E_I(Q_i, C_j) \quad (5.10)$$

Se dirá que un ítem evalúa un concepto cuando, o bien lo hace directamente, o bien indirectamente. Así, todos los ítems que proporcionan evidencias sobre un determinado concepto forman su banco de ítems.

En la figura 5.2, el ítem Q_6 evalúa directamente al concepto C_{311} . Ese ítem también suministra evidencias sobre el conocimiento del alumno en el concepto que precede a C_{311} , es decir, en C_{31} . Aplicando el mismo razonamiento, el ítem también aporta evidencias sobre el padre de éste, es decir, sobre C_3 . Por último, como se ha mencionado, la asignatura completa se considera un concepto que incluye a todos sus conceptos hijos, por consiguiente Q_6 también proporciona evidencias sobre el conocimiento del alumno global de la asignatura. Los ítems podrán evaluar directamente conceptos hoja, intermedios, o incluso a la raíz de la jerarquía (la propia asignatura). Sin embargo, cuando un ítem evalúa directamente un concepto intermedio, no se puede inferir directamente que también evalúe a sus descendientes.

En la figura 5.3 se han representado gráficamente todas las CCO asociadas a un ítem de opción múltiple con tres opciones de respuesta. Como se puede apreciar, el ítem Q_4 está asociado al concepto C_{1231} , y por lo tanto, lo evalúa directamente. Debajo de él se han representado las CCO correspondientes a las opciones de respuestas del ítem, según el nivel de conocimiento del alumno en ese concepto. La primera de ellas corresponde a la opción correcta, y las tres restantes corresponden a las dos respuestas erróneas y a la respuesta en blanco, respectivamente. Asimismo, Q_4 evalúa de forma indirecta a los conceptos Q_{123} , Q_{12} , Q_1 y a la asignatura completa. Por este motivo, debajo de cada uno de esos conceptos, se han representado las CCO correspondientes a las respuestas del ítem, en función del conocimiento del alumno en cada uno de esos conceptos.

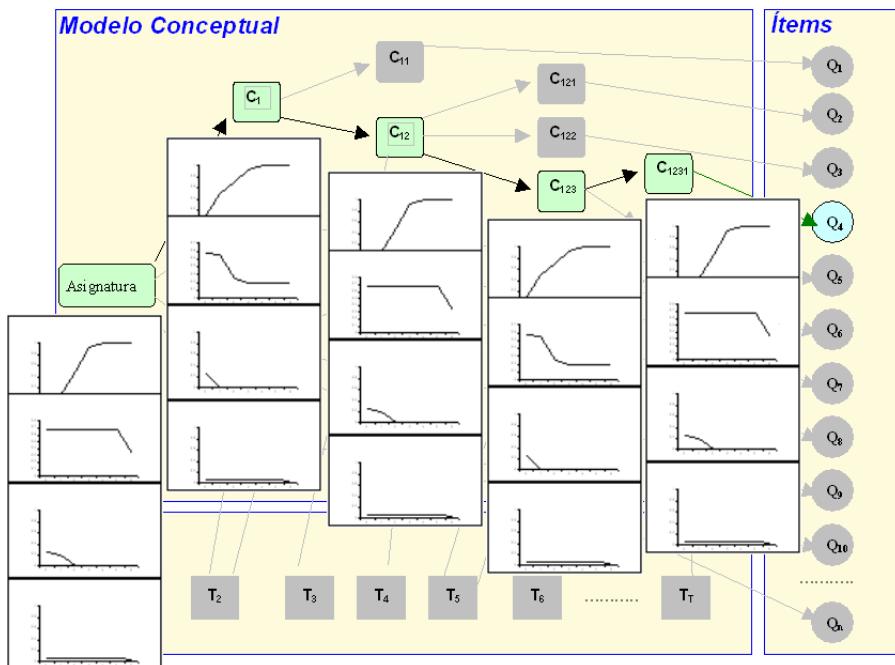


Figura 5.3: CCO de un ítem de opción múltiple.

Según la TRI, cada ítem permite evaluar un único rasgo latente, y esta relación viene determinada por la CCI, en caso de un modelo de respuesta dicotómico, o bien por las CCR, cuando se aplica el politómico. Extrapolando esta característica al modelo propuesto, si se considera cada concepto como un rasgo latente susceptible de ser evaluado, cada ítem deberá tener asociado un conjunto de CCR por cada concepto que evalúe (en el caso de un modelo dicotómico, una CCI por cada concepto). Sea n la profundidad del concepto C_j en el árbol curricular (modelo conceptual). Si el ítem Q_i evalúa directamente al concepto C_j , para ese ítem se definirán un número de CCR igual al producto del número de sus posibles patrones de respuestas multiplicado por n . Esto es, un conjunto de CCR (una por cada patrón de respuesta) por cada concepto evaluado directa o indirectamente por el ítem Q_i .

Para el funcionamiento adecuado del modelo de diagnóstico, es recomendable que el número de ítems de cada concepto esté balanceado, y tenga por lo tanto, un número similar de ítems. Esto contribuirá a mejorar la precisión de los criterios de selección y finalización adaptativos. Es responsabilidad del profesor, asegurarse de ello, de tal forma que el diagnóstico se lleve a cabo de forma correcta. Asimismo, tal y como se ha mencionado con anterioridad, un banco será mejor cuanta mayor variedad de ítems tenga, desde el punto de vista de las propiedades psicométricas de éstos.

Por último, el profesor determina el número de niveles de conocimiento de las curvas características, en función del nivel de detalle requerido en la evaluación. La única restricción que se impone, es el número de niveles de las curvas, que debe ser el mismo para todos los conceptos del currículo de una misma asignatura. Recuérdese que, en virtud del modelo de respuesta discreto que se ha definido en el capítulo anterior, el número de niveles de conocimiento deberá ser un número natural mayor que uno.

Los tests

En el modelo propuesto, la evaluación o diagnóstico del conocimiento de los alumnos se lleva a cabo mediante tests. Un test en sí no es más que una especificación basada en un conjunto de parámetros, a partir de la cual se generarán sesiones de evaluación. Su objetivo final es obtener una estimación del nivel de conocimiento del alumno en uno o varios de los conceptos del currículo.

Estableciendo una analogía con los STI, Chua Abdullah (2003) señala las tres preguntas fundamentales a las que debe responder un sistema de evaluación efectivo; no obstante, desde la perspectiva de la evaluación adaptativa, puede añadirse una más:

1. *¿Qué evaluar?*, esto es, sobre qué elementos del modelo conceptual se diagnosticará el conocimiento de los alumnos.
2. *¿A quién evaluar?*, es decir, cómo es el examinando. Esta información estará contenida en su modelo de usuario.
3. *¿Cómo evaluar?*, es decir: *a)* Qué criterio va a utilizarse, esto es, cómo va a ser inferida la calificación del alumno a partir de su actuación en el test. *b)* El nivel de detalle, es decir, en cuántos niveles de conocimiento va a ser evaluado. *c)* El ámbito, es decir, ¿afecta la evaluación únicamente a los conceptos indicados por el profesor, o va a afectar a alguno más? Y por último, *d)* *¿Cómo serán secuenciados los elementos de evaluación (los ítems)?*, es decir, qué criterio de selección de ítems se va a utilizar.
4. *¿Cuándo finalizar la evaluación?*, ya que en los criterios de evaluación adaptativos es necesario determinar a priori cuando se considerará que la estimación del conocimiento del alumno es suficientemente precisa.

En el modelo presentado, las respuestas a las preguntas anteriores se materializan en parámetros de configuración del test.

Los tests se definen en función de los conceptos que se desean evaluar. Como efecto colateral, y en virtud de la estructura del modelo conceptual y de las relaciones existentes entre ítems y conceptos, un test podrá evaluar también a conceptos diferentes de aquéllos indicados por el profesor a través de los parámetros de configuración del test. Para comprender mejor esta característica del modelo, se procederá a definir formalmente el siguiente conjunto de relaciones de evaluación entre tests y conceptos:

Definición 5.8 (Evaluación directa de un test sobre un concepto). Sea T_s un test de evaluación de una determinada asignatura ($T_s \in \Theta$), y C_j un concepto de esa asignatura ($C_j \in \Omega$), se define la función de evaluación directa de un test sobre un concepto $\Phi_D : \Theta \times \Omega \rightarrow \{0, 1\}$, de la siguiente forma:

$$\Phi_D(T_s, C_j) = \begin{cases} 1 & \text{si } C_j \text{ es uno de los temas seleccionados} \\ & \text{por el profesor para formar parte del test } T_s \\ 0 & \text{en otro caso} \end{cases} \quad (5.11)$$

Por ejemplo, en el módulo experto descrito en la figura 5.2, el test T_3 evalúa directamente al concepto C_{12} . En la figura, esta relación viene descrita por una línea que une el test con los conceptos a los que evalúa directamente.

Axioma 5.4. En un test se podrán evaluar directamente y de forma simultánea tantos conceptos como se desee, con la única restricción de que entre ellos no debe existir ninguna relación de ascendencia ni de descendencia; es decir, no debe existir ningún camino en la red semántica del modelo conceptual entre ellos. Sea el test T_s , que evalúa los conceptos C_j y C_k , esto se puede expresar formalmente según la siguiente fórmula:

$$\forall C_j, C_k \in \Omega, \quad \Phi_D(T_s, C_j) = 1 \wedge \Phi_D(T_s, C_k) = 1 \implies \\ C_j \notin \wp^n(C_k), \quad \forall n > 0 \quad \wedge \quad C_k \notin \wp^m(C_j), \quad \forall m > 0 \quad (5.12)$$

La restricción descrita en la fórmula 5.12 debe satisfacerse para todos los conceptos (evaluados directamente) en el test, dos a dos. Es importante destacar que no existe una relación directa entre tests e ítems. La relación entre ambos conjuntos se establece a través de los conceptos del modelo conceptual.

Definición 5.9 (Evaluación indirecta descendente (o hacia abajo) de un test sobre un concepto). Sean T_s y C_j , se define la función de evaluación indirecta descendente de un test sobre un concepto $\Phi_{I\downarrow} : \Theta \times \Omega \rightarrow \{0, 1\}$ de la siguiente forma:

$$\Phi_{I\downarrow}(T_s, C_j) = \begin{cases} 1 & \text{si } \exists C_h, \quad C_h \in \Omega, \quad \Phi_D(T_s, C_h) = 1 \wedge C_j \in \wp^n(C_h), \quad n > 0 \\ 0 & \text{en otro caso} \end{cases} \quad (5.13)$$

Consecuentemente, un test evaluará indirectamente hacia abajo a todos aquellos conceptos que sean descendientes de aquéllos directamente evaluados en el test. Por consiguiente, el test T_3 de la figura 5.2 evalúa indirectamente hacia abajo a C_{121} , C_{122} , C_{123} , C_{1231} y C_{1232} . Así, para un alumno que realice el test T_3 , a través de este modelo, se podrá diagnosticar, como efecto colateral, su nivel de conocimiento en todos esos conceptos, de forma simultánea.

Definición 5.10 (Evaluación indirecta ascendente (o hacia arriba) de un test sobre un concepto). Sean T_s y C_j , se define la función de evaluación indirecta ascendente de un test sobre un concepto $\Phi_{I\uparrow} : \Theta \times \Omega \rightarrow \{0, 1\}$ de la siguiente forma:

$$\Phi_{I\uparrow}(T_s, C_j) = \begin{cases} 1 & \text{si } \exists C_h, \quad C_h \in \Omega, \quad \Phi_D(T_s, C_h) = 1 \wedge C_h \in \wp^n(C_j), \quad n > 0 \\ 0 & \text{en otro caso} \end{cases} \quad (5.14)$$

Un test evaluará indirectamente hacia arriba a todos aquellos conceptos que sean ascendentes de aquéllos directamente evaluados en el test. Por ejemplo, el test T_3 de la figura 5.2 evalúa indirectamente hacia arriba a los conceptos C_1 y la asignatura completa, simultáneamente.

Definición 5.11 (Evaluación indirecta de un test sobre un concepto). Sean T_s y C_j , se define, generalizando las dos funciones anteriores, la función de evaluación indirecta de un test sobre un concepto $\Phi_I : \Theta \times \Omega \rightarrow \{0, 1\}$ de la siguiente forma:

$$\Phi_I(T_s, C_j) = \Phi_{I\downarrow}(T_s, C_j) + \Phi_{I\uparrow}(T_s, C_j) \quad (5.15)$$

Así, un test evaluará indirectamente a un concepto, si lo evalúa indirectamente hacia arriba o hacia abajo.

Definición 5.12 (Evaluación de un test sobre un concepto). Generalizando aún más, es posible definir la función de evaluación de un test sobre un concepto $\Phi : \Theta \times \Omega \rightarrow \{0, 1\}$ de la siguiente forma:

$$\Phi(T_s, C_j) = \Phi_D(T_s, C_j) + \Phi_I(T_s, C_j) \quad (5.16)$$

Un test evaluará a un concepto cuando lo haga de cualquiera de las tres formas anteriormente vistas. Esto es, directamente, indirectamente hacia abajo o indirectamente hacia arriba.

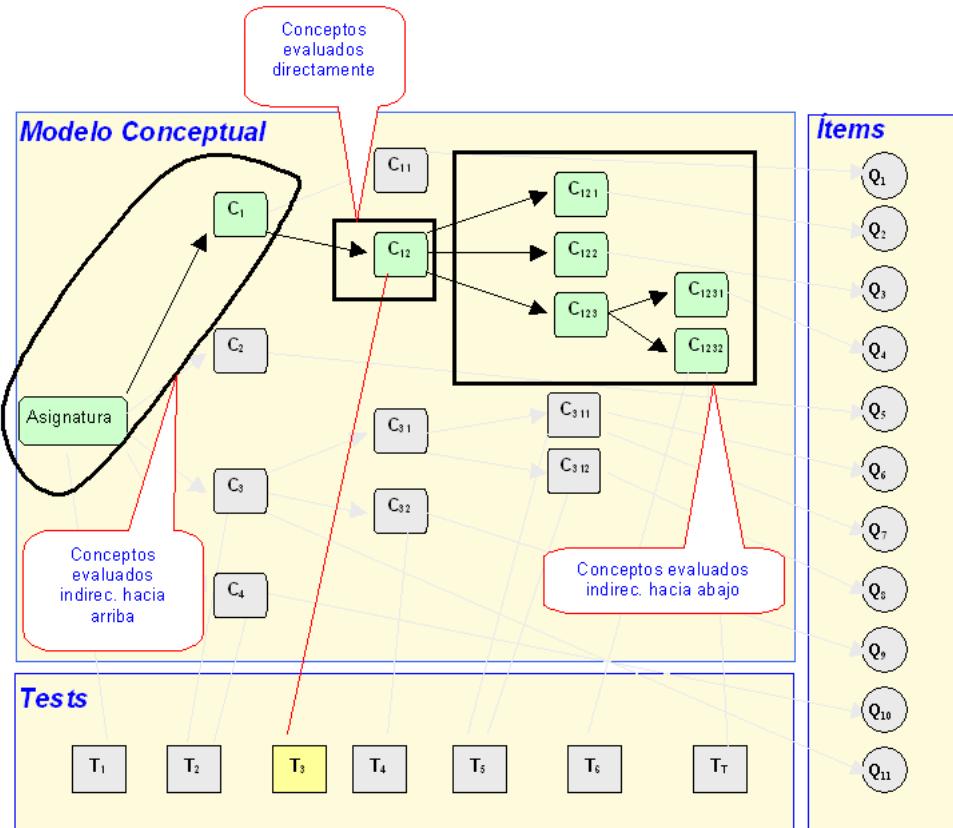


Figura 5.4: Conjuntos de conceptos evaluados por el test T_3 , en función del tipo de evaluación.

En la figura 5.4 se han representado las relaciones de evaluación entre tests y conceptos. Más concretamente, se han señalado los diferentes subconjuntos de conceptos que se evaluarán en el test T_3 , en función de los diferentes modos de evaluación.

Este modelo de diagnóstico permite administrar tests sobre múltiples conceptos de forma simultánea, entre los que existe una relación. En realidad no se mide el nivel de conocimiento en conceptos diferentes, sino en aquéllos incluidos unos dentro de otros, en virtud de las relaciones de agregación que los unen. Para entender mejor esto, puede tomarse el ejemplo descrito en (Smits et al., 2003), en donde se propone la siguiente operación matemática: $3 \cdot (4 + 5)$. El examinando, para poder calcular el resultado de ese ítem, tiene que saber

resolver operaciones algebraicas. Pero el saber operaciones algebraicas, para este ítem, supone saber aplicar las dos siguientes operaciones: $4 + 5$ y $3 \cdot 9$. Según este modelo, para este ejemplo, el concepto "*operaciones algebraicas*" se podría dividir en los concepto "*suma*" y "*multiplicación*", y por tanto, el ítem estaría asociado al concepto "*operaciones algebraicas*".

Cuando se crea un test en este modelo, se debe suministrar la siguiente información: (1) Conceptos evaluados directamente; (2) si desea que haya evaluación indirecta de algún tipo; (3) el criterio de selección de ítems que se utilizará; (4) cómo se inicializará el modelo del alumno; (5) cuándo debe finalizar el test; (6) cómo se evaluará al estudiante; y por último, (7) el número de niveles de conocimiento en los que se le podrá clasificar. En la sección 5.3, se explicarán detalladamente las diversas alternativas que ofrece este modelo de diagnóstico para los parámetros anteriores.

Relación entre el número de niveles de conocimiento del modelo conceptual y del test

El número de niveles de conocimiento con el que se puede configurar un test, está limitado por el número de niveles de conocimiento global de la asignatura. Este último valor lo establece el profesor antes de que se lleve a cabo el proceso de calibración de las curvas características de los ítems de la asignatura. Por consiguiente, si K es el número de niveles de conocimiento del currículo, todas las curvas características de los ítems (CCI, CCO y CCR) serán vectores de tamaño igual a K .

De esta forma, a través de un test T_s se podrá evaluar al alumno en un número de niveles de conocimiento K^T , entre 2 y el número de niveles de conocimiento de la asignatura, esto es, $2 \leq K^T \leq K$. En caso de que el número de niveles del test sea menor que el del currículo, las curvas características de los ítems implicados en la sesión de evaluación, sufrirán temporalmente una transformación para adecuarlas a los requisitos del test. Por este motivo, se establecerá una correspondencia entre las probabilidades de la curva característica original correspondiente (con un rango formado por K niveles de conocimiento), y las probabilidades de la curva característica transformada con K^T niveles de conocimiento, $K^T < K$. En la ecuación 5.17 se muestra formalmente, cómo se calculan los valores del vector correspondiente a una CCO transformada para su uso en un test basado en K_T niveles de conocimiento, a partir de la CCO original con K niveles de conocimiento:

$$\forall k_T, \quad 0 \leq k_T < K_T \\ P_{ij}^T(S(r_j) = 1 | \theta = k^T) = \begin{cases} \sum_{k=k^T\delta}^{(k^T+1)\delta-1} P_{ij}(S(r_j) = 1 | \theta = k) & \text{si } 0 \leq k^T < (K^T - 1) \\ \sum_{k=k^T\delta}^{(k^T+1)\delta-1+\rho} P_{ij}(S(r_j) = 1 | \theta = k) & \text{si } k^T = (K^T - 1) \end{cases} \quad (5.17)$$

donde $P_{ij}^T(S(r_j) = 1 | \theta)$ es la CCO transformada para el test, y $P_{ij}(S(r_j) = 1 | \theta)$ la original. Asimismo, δ es el cociente entero de la división entre el número de niveles de conocimiento del currículo y el del test, es decir: $\delta = \lceil \frac{K}{K^T} \rceil$; y ρ es el resto de la división entera anterior, esto es: $\rho = K - K^T\delta$.

Cuando K^T es un divisor de K , ρ es igual a cero. En este caso, la ecuación 5.17 podría simplificarse, quedando sólo la primera de las dos expresiones a la derecha del igual. Aunque esto es lo deseable, pueden existir situaciones en las que K^T no sea un divisor de K . En estas ocasiones, los valores del vector de probabilidades de la CCO no se calcularán a partir de la suma del mismo número de sumandos (probabilidades de la CCO original). Como

se puede apreciar en la ecuación 5.17, el último valor del vector de la CCO transformada será igual a la suma de un número de mayor de sumandos que el resto de valores.

Por ejemplo, si las CCO del currículo tienen un dominio de 12 niveles de conocimiento, y el test se ha definido para 6 de ellos, la correspondencia entre las curvas originales y las transformadas es muy sencilla. No hay más que sumar dos a dos las probabilidades asociadas a los niveles de conocimiento de la curva original, dando lugar al valor de probabilidad asociado a un nivel de la curva transformada. Es decir, la suma de las probabilidades de los niveles 0 y 1 se harán corresponder con la probabilidad del nivel 0 de la curva transformada, la suma de los niveles 2 y 3 con la probabilidad 1 de la curva transformada, y así sucesivamente. Por el contrario, si K^T fuese, por ejemplo 5, las probabilidades de los dos últimos valores de la curva transformada tendrían que ser iguales a la suma de tres de las probabilidades de la curva original, resultando una transformación menos equitativa.

En el modelo de diagnóstico propuesto, como se verá más adelante en este capítulo, las distribuciones del conocimiento del alumno, tras finalizar un test, se almacena de forma permanente. Así, si un estudiante realizó anteriormente un test T_1 que involucraba a un determinado concepto C_j , como resultado de ese test, se habrá inferido la distribución de su conocimiento, $P(\theta_j|\vec{\mathbf{u}}_i)$, en C_j . Por consiguiente, si ese alumno inicia posteriormente otro test T_2 que también involucra a C_j , este test podría utilizar como distribución de conocimiento inicial en ese concepto la inferida en el test T_1 . Supóngase ahora que T_1 evaluó al alumno en K niveles de conocimiento, pero que por el contrario, T_2 va a evaluar al alumno en K^T niveles de conocimiento, tal que: $K_T < K^T$. En este caso, habría que hacer una transformación análoga a la que se realiza con las CCO, para adecuar la distribución del conocimiento del alumno al número de niveles de conocimiento del nuevo test, antes que éste comience. Esta transformación puede expresarse de la siguiente forma:

$$\forall k_T, \quad 0 \leq k_T < K_T$$

$$P^T(\theta_j = k^T | \vec{\mathbf{u}}_i) = \begin{cases} \sum_{k=k^T\delta}^{(k^T+1)\delta-1} P(\theta_j = k | \vec{\mathbf{u}}_i) & \text{si } 0 \leq k^T < (K^T - 1) \\ \sum_{k=k^T\delta}^{(k^T+1)\delta-1+\rho} P(\theta_j = k | \vec{\mathbf{u}}_i) & \text{si } k^T = (K^T - 1) \end{cases} \quad (5.18)$$

donde $P^T(\theta_j | \vec{\mathbf{u}}_i)$ es la distribución transformada del conocimiento del alumno en el concepto C_j .

5.2.2. El modelo del alumno

En un sistema para el diagnóstico del conocimiento, la relevancia del modelo del alumno es alta. Esto es debido a que una vez llevado a cabo el proceso de estimación del conocimiento del alumno en el sistema, se realiza una actualización de su modelo en el STI a partir de su modelo de usuario del sistema de diagnóstico, y esto permite mejorar el proceso de aprendizaje.

En la sección 3.2 se describieron los distintos tipos de modelos del alumno. En esta propuesta, se usa un modelo de superposición sobre el mapa conceptual del módulo experto. En él, por cada concepto C_i , se almacena una distribución discreta de probabilidades $P(\theta_i | \vec{\mathbf{u}}_n)$, que representa el conocimiento del alumno (θ_i) en ese concepto C_i . El rango de esta distribución son los niveles de conocimiento en los que se evalúa el test, y el dominio la probabilidad de que el conocimiento del alumno en C_i sea el nivel correspondiente. Asimismo, la suma de las probabilidades de esa distribución debe ser igual a uno:

$$\sum_{k=0}^{K-1} P(\theta_i = k | \vec{\mathbf{u}}_n) = 1 \quad (5.19)$$

Esta distribución de probabilidades es inferida a partir de las respuestas del alumno a los n ítems del test que evalúan ese concepto, y a los cuales habrá respondido con el patrón de respuestas $\vec{\mathbf{u}}_n = \{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_n\}$. Su número de niveles de conocimiento dependerá del grado de detalle con el que el profesor correspondiente quiera evaluar a sus examinandos.

Este modelo almacena también una traza de cada sesión de evaluación. Por cada individuo, se almacenan los tests que ha realizado y la composición de éstos. La información de cada sesión está formada por los ítems que fueron mostrados, el orden en que se presentaron, y qué patrón de respuestas se dio a cada uno de ellos. Esta información tendrá especial relevancia para la calibración de los ítems, como se verá más adelante en este capítulo. Asimismo también puede ser útil si se desea implementar el modelo de diagnóstico en un sistema con modelos del alumno abiertos, ya que permitiría a los estudiantes consultar sus resultados en los tests que han completado.

5.3. Funcionamiento del modelo

Si este modelo de diagnóstico se utiliza en un STI, cuando un profesor construya un curso, deberá a su vez elaborar el correspondiente modelo conceptual del módulo experto, en función de los conceptos estudiados en el STI. Deberá diseñar y añadir ítems a ese módulo experto, asociándolos con los conceptos correspondiente. Finalmente, especificará cuales serán las características de las sesiones de evaluación, mediante la construcción de tests.

El funcionamiento del modelo, esto es, el procedimiento de diagnóstico, se ha representado en la figura 5.5, y supone la administración de un test adaptativo al alumno. Una vez seleccionado el test que se desea suministrar, el procedimiento de diagnóstico que se lleva a cabo se compone de los siguientes pasos (Guzmán y Conejo, 2004a):

1. *Selección de los ítems disponibles para el test:* Se elige una colección de ítems Ψ_s , $\Psi_s \subseteq \Pi$, igual a la unión de los bancos de ítems de todos los conceptos evaluados en el test T_s . Formalmente, los ítems que constituyen el banco del test T_s deben cumplir la siguiente condición:

$$\forall Q_i, \quad Q_i \in \Pi, \quad Q_i \in \Psi_s \Leftrightarrow \exists C_j, \quad C_j \in \Omega, \quad E(Q_i, C_j) = 1 \quad \wedge \quad \Phi(T_s, C_j) = 1$$

2. *Creación e inicialización del modelo del alumno:* En función de los conceptos implicados en el test, se crean tantas distribuciones de probabilidades del conocimiento del alumno, como conceptos se evalúen. Más adelante, se verá cómo éstas son inicializadas.
3. *Aplicación del test adaptativo:* Finalmente, al alumno se le administra el test.

La última fase de aplicación del test adaptativo no es más que una generalización del algoritmo de funcionamiento de un TAI (figura 2.7), llevada a cabo para adecuarlo al modelo de diagnóstico propuesto. A continuación se enumeran las fases de las que se compone este nuevo algoritmo:

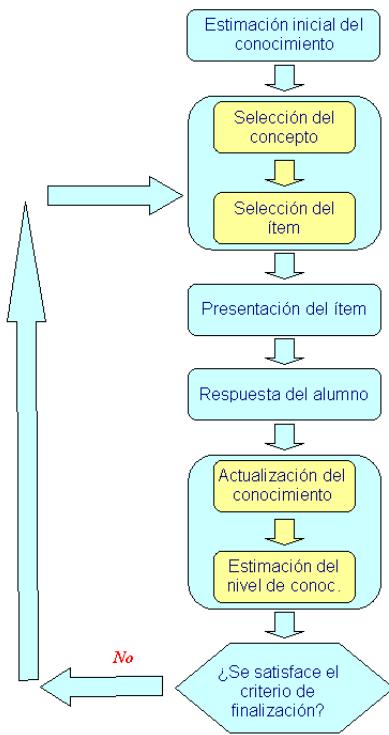


Figura 5.5: Esquema de funcionamiento del modelo de diagnóstico.

1. A partir del conjunto Ψ_s , se selecciona aquel ítem que contribuye, en mayor medida, a una mejor estimación del conocimiento del alumno. Como se verá a continuación, esta selección supone una doble elección. Por una lado, se escoge aquel concepto del cual se tiene información más imprecisa, en cuanto a conocimiento del alumno se refiere. Posteriormente, se elige, de entre los ítems que evalúan a ese concepto, el que es más informativo, es decir, será aquél tras cuya administración, la estimación del conocimiento es más precisa. Para la selección del ítem, se utiliza, como información de entrada, las distribuciones probabilísticas del conocimiento del alumno, las cuales se irán estimando a partir de las evidencias suministradas por los patrones de respuesta que el examinando haya dado a los ítems administrados hasta ese momento.
2. El ítem seleccionado se retira del banco de ítems del test.
3. El ítem es presentado al alumno.
4. En función del patrón de respuestas proporcionado por el alumno, se actualiza su distribución del conocimiento en los conceptos correspondientes, según indique el criterio de evaluación del test.
5. Se estima el nivel de conocimiento del alumno en las distribuciones recién actualizadas.
6. Los pasos del 1 al 5 se repiten hasta que la condición de terminación del test se satisfaga en los conceptos correspondientes, según determine su criterio de finalización.

5.3.1. La estimación del conocimiento del alumno

Durante la administración de un test, el conocimiento del alumno se estima cada vez que éste responde a un determinado ítem. La actualización de la distribución del conocimiento del examinando se lleva a cabo utilizando una adaptación del método bayesiano propuesto por Owen (1969, 1975), que fue descrito en la sección 2.9.1.

En este modelo, se han definido diversos modos de evaluación en función de los conceptos que se ven implicados en el proceso de inferencia del conocimiento. Consecuentemente, en función del ámbito de aplicación del proceso de diagnóstico, se pueden definir las siguientes formas de evaluación:

- *Evaluación agregada:* Se aplica cuando sólo se desea inferir el conocimiento del alumno en conceptos directamente evaluados en el test T_s , esto es:

$$\forall C_t, \quad C_t \in \Omega, \quad \Phi_D(T_s, C_t) = 1$$

Por consiguiente, una vez que el alumno responde al i -ésimo ítem Q^i , sus distribuciones de conocimiento se actualizan aplicando la siguiente fórmula:

$$P(\theta_t | \vec{u}_1, \dots, \vec{u}_i) = \begin{cases} \|P(\theta_t | \vec{u}_1, \dots, \vec{u}_{i-1}) P_{i\vec{u}_i}(\vec{u}_i | \theta_t)\| & \text{si } E(Q^i, C_t) = 1 \wedge \Phi_D(T_s, C_t) = 1 \\ P(\theta_t | \vec{u}_1, \dots, \vec{u}_{i-1}) & \text{en otro caso} \end{cases} \quad (5.20)$$

donde \vec{u}_i representa el vector con el patrón de respuesta que el alumno ha seleccionado en el ítem i -ésimo del test, y θ_t su nivel de conocimiento en el concepto C_t . La CCR de ese patrón de respuesta para el concepto C_t es $P_{i\vec{u}_i}(\vec{u}_i | \theta_t)$. $P(\theta_t | \vec{u}_1, \dots, \vec{u}_{i-1})$ es la distribución a priori del conocimiento en C_t , esto es, la distribución antes de que el alumno responda al i -ésimo ítem. La doble línea vertical indica que después calcular el producto, la distribución resultante debe ser normalizada para la suma de sus valores sea igual a uno.

El ítem mostrado en i -ésima posición, suministra por tanto, una evidencia que permite inferir el conocimiento del alumno en el concepto C_t . Así, cada vez que éste responde a un determinado ítem, su patrón de respuesta es capturado por el módulo de presentación, que le transfiere este resultado al módulo de evaluación, el cual sólo actualiza la distribución de conocimiento en el concepto correspondiente.

En la figura 5.6 se ha representado el modelo del alumno para el test T_2 cuando éste se configura con el modo de evaluación agregada. Las distribuciones del conocimiento del examinando se han representado como etiquetas sobre el concepto correspondiente. Como se puede apreciar, en este caso, el modelo del alumno se compone de sólo dos distribuciones de conocimiento: la del concepto C_3 ($P(\theta_3 | \vec{u}_i)$), y la de C_4 ($P(\theta_4 | \vec{u}_i)$).

- *Evaluación completa:* Gracias a la estructura del modelo conceptual, y en virtud de las relaciones que se establecen entre los conceptos, y entre éstos y los ítems, es posible inferir, durante una misma sesión de evaluación, el conocimiento del alumno en algunos de los conceptos descendientes de aquéllos directamente evaluados en el test. Es decir, una vez que el sujeto responde a un ítem, esta respuesta representa una evidencia sobre su conocimiento, no sólo en el concepto directamente evaluado en el test, sino también en todos aquéllos descendientes de éste. Por consiguiente, esta evidencia se propaga desde el concepto objeto del test, hacia todos aquéllos que se encuentren en el camino

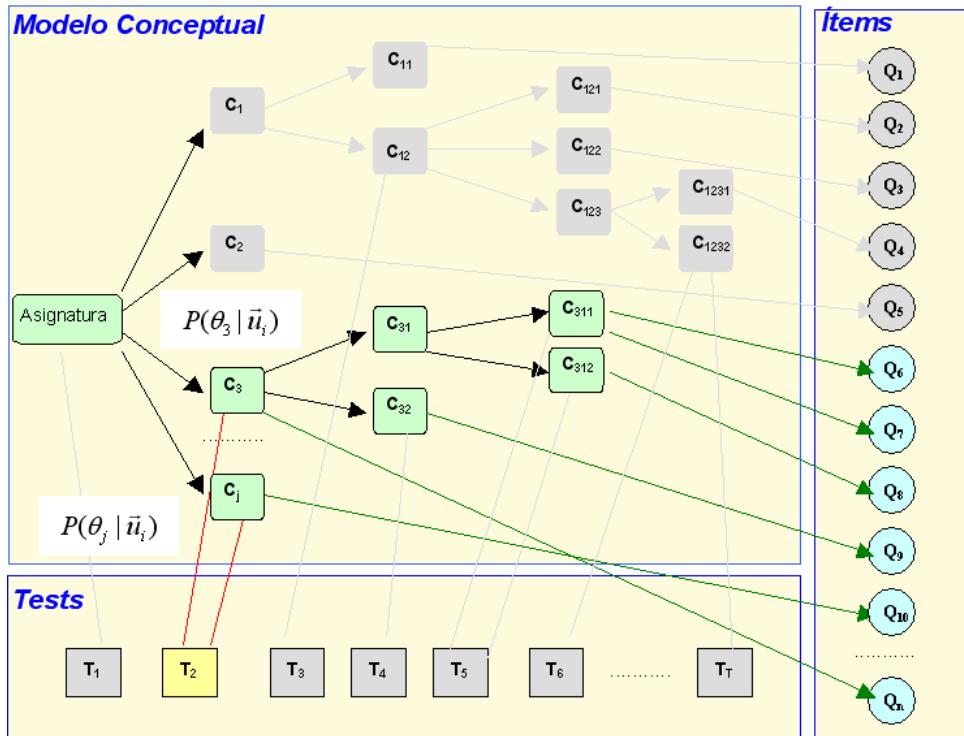


Figura 5.6: Modelo del alumno para evaluación agregada.

entre éste y el concepto evaluado directamente por el ítem, ambos inclusive. Así, sea C_t el concepto directamente evaluado en el test, y C_r aquél directamente evaluado por el ítem, tal que $C_r \in \wp^n(C_t)$ con $n > 0$, el conocimiento se actualizará en todo aquel concepto C_s descendiente de C_t , es decir, $C_t \in \wp^{n'}(C_s)$, con $n' \leq n$.

Generalizando, para este tipo de evaluación, el proceso de actualización del conocimiento del alumno se puede expresar formalmente como indica la siguiente ecuación:

$$P(\theta_t | \vec{u_1}, \dots, \vec{u_i}) = \begin{cases} \|P(\theta_t | \vec{u_1}, \dots, \vec{u_{i-1}}) P_{i\vec{u_i}}(\vec{u_i} | \theta_t)\| & \text{si } E(Q^i, C_t) = 1 \wedge \\ & \Phi_D(T_s, C_t) + \Phi_{I\downarrow}(T_s, C_t) = 1 \\ P(\theta_t | \vec{u_1}, \dots, \vec{u_{i-1}}) & \text{en otro caso} \end{cases} \quad (5.21)$$

La figura 5.7 representa el modelo del alumno para el test T_2 , cuando éste se configura con el modo de evaluación completa y, por consiguiente, se compone de seis distribuciones de conocimiento: las de los conceptos C_3 , C_4 , C_{31} , C_{32} , C_{311} y C_{312} , es decir, $P(\theta_3 | \vec{u_i})$, $P(\theta_4 | \vec{u_i})$, $P(\theta_{31} | \vec{u_i})$, $P(\theta_{32} | \vec{u_i})$, $P(\theta_{311} | \vec{u_i})$ y $P(\theta_{312} | \vec{u_i})$.

- *Evaluación completa con retropropagación (o con propagación hacia atrás):* Por analogía con el modo de evaluación anterior, la actualización del conocimiento del alumno puede extenderse para afectar también a aquellos conceptos ascendientes de los directamente evaluados en el test. Así, en este modo de evaluación, la evidencia proporcionada por la respuesta al ítem se propaga no sólo hacia abajo, sino también hacia arriba

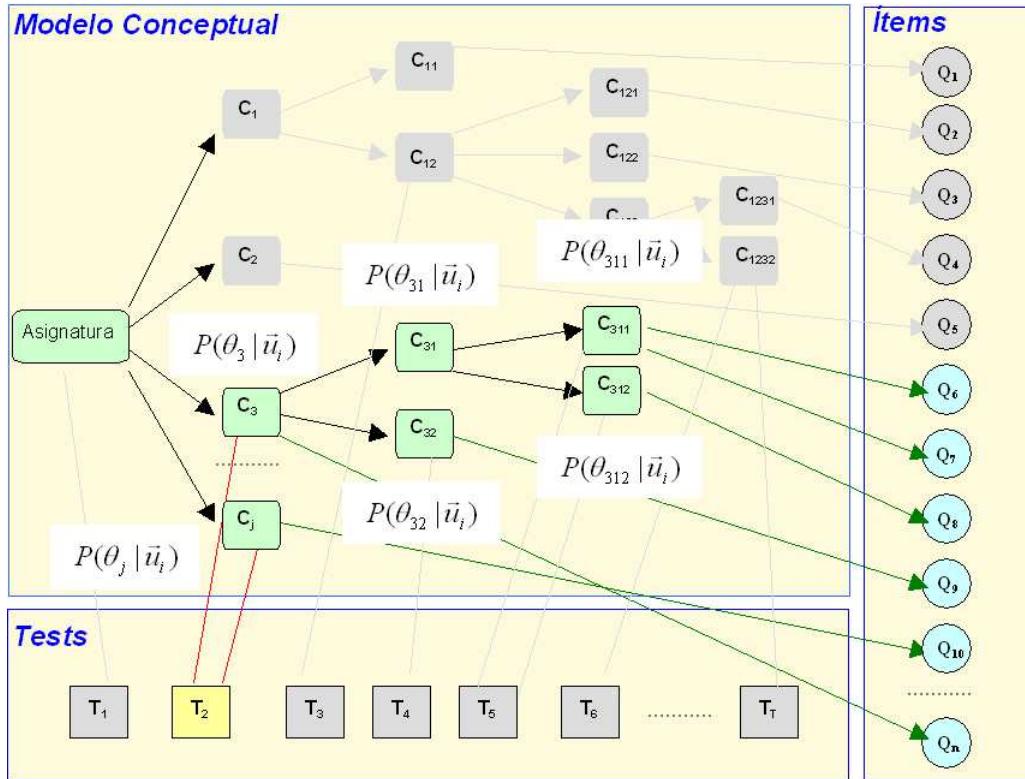


Figura 5.7: Modelo del alumno para evaluación completa.

en el árbol curricular. Esto es, hacia todos aquellos conceptos que preceden al evaluado directamente por el ítem. Como consecuencia, la actualización del conocimiento para esta modalidad de evaluación se realizará como se muestra a continuación:

$$P(\theta_t | \vec{u_1}, \dots, \vec{u_i}) = \begin{cases} \|P(\theta_t | \vec{u_1}, \dots, \vec{u_{i-1}}) P_{\vec{u_i}}(\vec{u_i} | \theta_t)\| & \text{si } E(Q^i, C_t) = 1 \wedge \Phi(T_s, C_t) = 1 \\ P(\theta_t | \vec{u_1}, \dots, \vec{u_{i-1}}) & \text{en otro caso} \end{cases} \quad (5.22)$$

Este último modo de evaluación es el más completo, puesto que implica al mayor número de conceptos posible. Asimismo, es necesario indicar que, en este caso, las estimaciones de los conceptos que preceden a aquéllos evaluados directamente en el test pueden estar sesgadas. Cuando se actualiza el conocimiento del alumno en los conceptos evaluados en el test indirectamente hacia abajo, se actualiza también en todos aquéllos situados al mismo nivel de la jerarquía, puesto que son nodos descendientes de aquéllos evaluados directamente en el test. Por el contrario, cuando se actualiza la distribución del conocimiento en los conceptos evaluados en el test indirectamente hacia arriba, sólo se obtienen evidencias de los descendientes de ese nodo por una rama del mapa conceptual, con lo que la información resultante es parcial, y por lo tanto, las estimaciones realizadas pueden presentar un sesgo considerable.

La figura 5.8 representa el modelo del alumno para el test T_2 , cuando éste se configura bajo el modo de evaluación completa con retropropagación. En esta ocasión, se com-

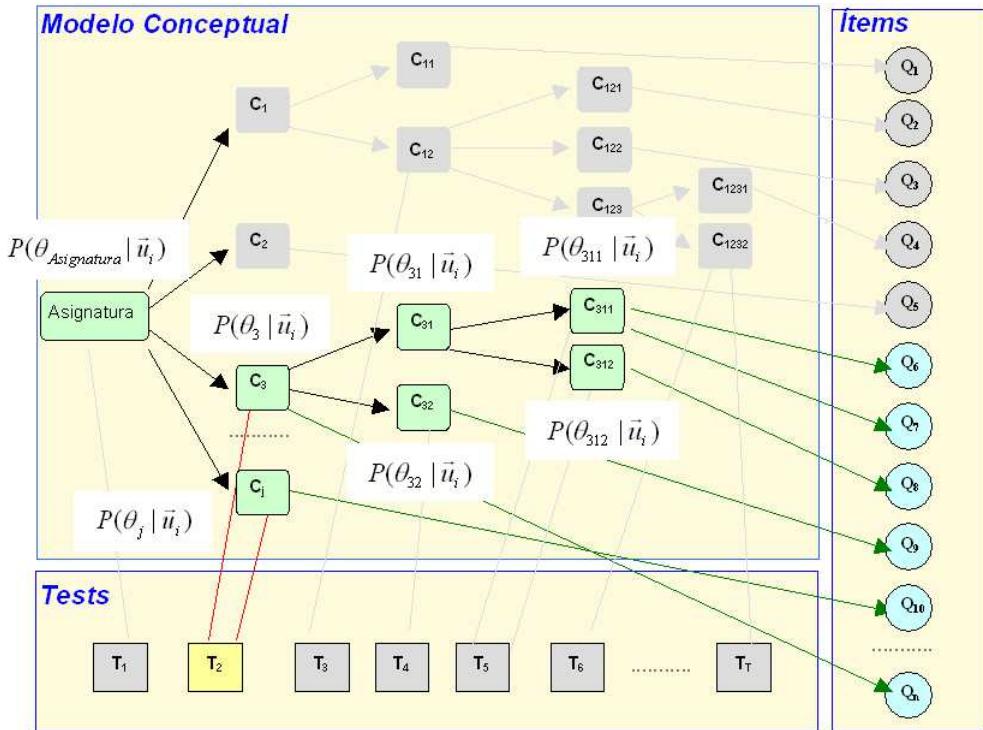


Figura 5.8: Modelo del alumno para evaluación completa con retropropagación.

pone de siete distribuciones de conocimiento: las de los conceptos $C_3, C_4, C_{31}, C_{32}, C_{311}, C_{312}$ y el concepto que representa a la asignatura, esto es, $P(\theta_3|\vec{u}_i)$, $P(\theta_4|\vec{u}_i)$, $P(\theta_{31}|\vec{u}_i)$, $P(\theta_{32}|\vec{u}_i)$, $P(\theta_{311}|\vec{u}_i)$, $P(\theta_{312}|\vec{u}_i)$ y $P(\theta_{Asignatura}|\vec{u}_i)$.

Este modo de evaluación puede ser útil como punto de partida de una estimación más precisa y de contenido balanceado. Por ejemplo, supóngase un STI en el que este modelo se utiliza como herramienta de diagnóstico. Considérese también que el mapa conceptual corresponde al de la figura 5.2, y que el STI va instruyendo al alumno sobre los diversos conceptos del currículo. Una vez finalizada la instrucción sobre, por ejemplo, el concepto C_{11} , el STI procederá a actualizar su modelo del alumno para que el planificador de instrucción decida la siguiente acción que debe llevarse a cabo. Este diagnóstico del conocimiento en C_{11} se hará mediante un test sobre ese concepto. Si ese test, tiene evaluación completa con retropropagación, se actualizarán también las distribuciones de conocimiento en los conceptos C_1 y en el que representa a la asignatura completa. En este punto, la evaluación de C_1 y del que representa a la asignatura será parcial. Posteriormente, después de que se lleve a cabo la instrucción sobre C_{12} , se realizará un test exclusivo para él. Este test tendrá también evaluación completa con retropropagación, y las distribuciones de conocimiento inicial para C_1 y la asignatura, serán las resultantes del test anterior. Por este motivo, tras realizar el test sobre C_{12} , la estimación en C_1 ya no será parcial. Como resultado, se evita tener que realizar un test adicional que diagnostique el nivel de conocimiento en C_1 . Igualmente, conforme se vayan realizando tests sobre los conceptos C_2, C_3 , etc., la estimación global de la asignatura será cada vez más precisa.

Inferencia del nivel estimado

Una vez actualizadas las distribuciones del conocimiento del alumno, se puede estimar su nivel empleando las dos formas utilizadas en los TAI; esto es:

- *Esperanza a posterior (EAP)*, donde el valor correspondiente al nivel de conocimiento es la media (o valor esperado) de la distribución de probabilidades. Formalmente, esto se puede expresar como sigue:

$$EAP(P(\theta_t|\vec{\mathbf{u}_n})) = \sum_{k=0}^{K-1} k P(\theta_t = k|\vec{\mathbf{u}_n}) \quad (5.23)$$

- *Máximo a posterior (MAP)*, donde el valor correspondiente al nivel de conocimiento es aquél con mayor probabilidad asignada, esto es, la moda de la distribución. Formalmente, esto se expresa de la siguiente manera:

$$MAP(P(\theta_t|\vec{\mathbf{u}_n})) = \max_{0 \leq k < K} P(\theta_t = k|\vec{\mathbf{u}_n}) \quad (5.24)$$

Por último, obsérvese que todas las estimaciones resultantes serán utilizada por los criterios adaptativos de selección cuyo cometido es determinar el siguiente ítem que debe administrarse en el test; y por los criterios de finalización, para comprobar si las estimaciones son lo suficientemente precisas.

Criterios de estimación heurísticos

Además de los presentados basados en la TRI, el modelo incluye dos criterios heurísticos, cuya finalidad es ser utilizados en tests convencionales con ítems sin calibrar. Son los siguientes:

- 1) *Criterio porcentual*: Los ítems se evalúan de forma dicotómica, esto es, como correctos o incorrectos. Una vez finalizado el test, la calificación que se asigna al examinando es el porcentaje de ítems respondidos correctamente, sobre el total de los realizados.
- 2) *Criterio por puntos*: En este caso, se asume que a cada ítem tiene una puntuación (por defecto igual a uno) si se responde de forma completamente correcta. La calificación del alumno se puede expresar directamente como la suma de todos los puntos que ha obtenido en los ítems que le han sido administrados; o bien como el porcentaje de puntos que ha obtenido, sobre el máximo alcanzable. Asimismo, en función del tipo de ítem, las respuestas parciales se podrán contabilizar como parte del valor global aportado por el ítem. Sea s_i la puntuación asignada al ítem i cuando éste es totalmente correcto. Para el resto de los casos, se asignan los siguientes valores parciales:
 - a) Ítems verdadero/falso: No permiten puntuaciones parciales. Se les asigna s_i si la respuesta es correcta, y cero en caso contrario.
 - b) Ítems de opción múltiple: Tampoco permiten puntuaciones parciales. Se les asigna s_i si la opción de respuestas seleccionada es la correcta, y cero en cualquier otro caso.

- c) Ítems de respuesta múltiple con opciones independientes: La puntuación obtenida por el alumno en el ítem será igual al producto del cociente entre la puntuación máxima s_i y total de opciones m_i , multiplicado por el número de las que son correctas. Se dice que una opción es correcta si lo es según el enunciado y el examinando la ha seleccionado, o bien si no lo es y no la ha elegido.
- d) Ítems de respuesta múltiple con opciones dependientes: En este caso, las respuestas (o respuestas) correctas son combinaciones de opciones. Si el alumno selecciona todos los elementos (y ninguno más) de una de ellas, recibirá la puntuación s_i . Asimismo, también se deja abierta al profesor la posibilidad de asignar puntuaciones parciales a combinaciones de opciones que no son totalmente correctas.
- e) Ítems de ordenación: Si el alumno ordena correctamente todos los elementos, se le asigna la puntuación s_i . Para los otros casos, se puntúan ordenaciones parcialmente correctas. Para ello, hay dos posibles estrategias:
 - *Orden relativo*, que supone contar el número de elementos que están en orden. Por ejemplo, sea un ítem en el que se muestran cinco números naturales (5, 4, 1, 3, 2), los cuales hay que ordenar de forma ascendente, una ordenación con un único error sería la siguiente: (1, 2, 3, 5, 4). Como se puede apreciar, en este caso, únicamente el 5 está mal ordenado. Todas aquellas ordenaciones con un único error podrían recibir una parte de la puntuación total del ítem. El mismo razonamiento podría seguirse con aquellas ordenaciones con dos errores, y así sucesivamente.
 - *Número de elementos contiguos* en la ordenación. Así, para el ítem anterior, la ordenación (1, 2, 4, 5, 3) tendría dos parejas de elementos contiguos, la pareja formada por el 1 y el 2, y la pareja formada por el 4 y el 5.
- f) Ítems de emparejamiento: Si el alumno enlaza correctamente todos los pares de elementos, se le asigna la puntuación s_i . En otro caso, se permite también puntuaciones parciales de las parejas correctamente relacionadas. Si cada columna tiene A elementos, la puntuación que obtendría el alumno sería igual a $\frac{s_i}{A}$ multiplicado por el número de parejas correctas.
- g) Ítems de asociación: Es un caso similar al anterior. La diferencia está en que aquí, la puntuación parcial se calcula dividiendo s_i entre el número total de emparejamientos correctos. La calificación del alumno en el ítem será igual al número de relaciones correctas multiplicado por el cociente anterior.

Obsérvese que este criterio por puntos puede incorporar aspectos más avanzados como el uso de ayudas junto con los enunciados de los ítems. En el capítulo siguiente se mostrará como, el sistema que implementa el modelo, permite incluirlas asociadas a los ítems. Por esta razón, en la implementación de este criterio de evaluación se permite penalizar el uso de estas ayudas.

Hay que volver a hacer hincapié en que estos últimos criterios de son heurísticos, y por lo tanto sólo es recomendable utilizarlos en aquellos casos en los que no se haya llevado a cabo la calibración de los ítems.

Estimación inicial del conocimiento

Al comienzo del proceso de diagnóstico, antes de que el alumno responda a ningún ítem, se asumen distribuciones a priori por cada concepto involucrado en la sesión de evaluación.

En este modelo, si no existe ninguna información adicional sobre el conocimiento del examinando en un determinado concepto, se asume una distribución constante, en la que todos los niveles son equiprobables. Si el STI suministra al modelo de diagnóstico información sobre cuánto sabe el alumno en ese concepto mediante un valor numérico, se construye una distribución de probabilidades normal centrada en ese valor. Por último, si el estudiante realizó algún test con anterioridad que involucrara ese concepto, el modelo permite utilizar la distribución de probabilidades resultante de esa sesión de evaluación previa.

5.3.2. Criterios para la selección de ítems

A diferencia de los modelos de TAI, este modelo de diagnóstico se caracteriza porque permite evaluar, en un mismo test, más de un concepto de forma simultánea. Por este motivo, en el modelo propuesto, el proceso de selección de ítems se compone de dos fases: *selección* del concepto, y de entre los ítems que evalúan a ese concepto, la *elección* de aquél que contribuye en mayor medida a obtener una estimación más precisa del conocimiento del alumno.

Como en cualquier modelo de TAI, el objetivo del proceso de selección seguirá siendo el minimizar el número de ítems necesario para estimar, de forma precisa, el conocimiento del examinando. Tiene por tanto, dos niveles de adaptación: Un primer nivel en el que se determina el concepto que debe evaluarse en función del estado de la estimación, y un segundo en el que se elige el ítem que será administrado.

A lo largo de esta sección se estudiarán los criterios de selección adaptativa que se proponen. Éstos se aplicarán de forma diferente, en función del modo de evaluación empleado.

Método bayesiano de la máxima precisión esperada

Este método fue introducido en la sección 2.9.1, donde se aplicó en la evaluación del conocimiento sobre un único concepto empleando ítems dicotómicos. En el modelo propuesto, permite llevar a cabo la selección de ítems en tests sobre múltiples conceptos. Además, se ha extendido para poder aplicarlo al modelo de respuesta politómico utilizado.

El objetivo, al igual que en su definición original, es seleccionar aquel ítem que minimiza la esperanza de la varianza de la distribución del conocimiento del alumno a posteriori. Supóngase que un sujeto está realizando un test que evalúa un conjunto φ de t conceptos $\varphi = \{C_1, C_2, \dots, C_t\}$, sea cual fuere el modo de evaluación, donde $\varphi \subseteq \Omega$. Considérese también que el alumno ha respondido previamente a $i - 1$ ítems, y que $\overrightarrow{\mathbf{u}_{i-1}} = \{\overrightarrow{u_1}, \overrightarrow{u_2} \dots \overrightarrow{u_{i-1}}\}$ es la matriz de los patrones de respuesta del alumno en cada ítem administrado. Para calcular cuál es el siguiente ítem Q_j que debe administrarse en i -ésima posición, por cada ítem del banco (que no haya sido todavía administrado), se calcula la esperanza de la varianza de la distribución del conocimiento a posteriori, suponiendo que el ítem elegido es Q_j . Se seleccionará aquel ítem con el que se obtenga el valor mínimo.

Formalmente, si se considera que cada ítem tiene $W + 1$ posibles patrones de respuestas $\{\overrightarrow{u_{i0}}, \overrightarrow{u_{i1}}, \dots, \overrightarrow{u_{iW}}\}$, el que debe administrarse a continuación es aquél que cumpla:

$$\min_{Q_j \in \Psi} \sum_{s=1}^t \sum_{w=0}^W \sigma^2[\rho_w(\theta_s | \overrightarrow{\mathbf{u}_{i-1}}, \overrightarrow{u_j})] v_{jsw} \quad (5.25)$$

donde

$$\rho_w(\theta_s | \overrightarrow{\mathbf{u}_{i-1}}, \overrightarrow{u_j}) = \begin{cases} \|P(\theta_s | \overrightarrow{\mathbf{u}_{i-1}}) P_{j\overrightarrow{u_w}}(\overrightarrow{u_w} | \theta_s)\| & \text{si } E(Q_j, C_s) = 1 \\ P(\theta_s | \overrightarrow{\mathbf{u}_{i-1}}) & \text{en otro caso} \end{cases} \quad (5.26)$$

y

$$v_{jsw} = \begin{cases} P(\theta_s | \overrightarrow{\mathbf{u}_{i-1}}) \cdot P_{j\overrightarrow{u_w}}(\overrightarrow{u_w} | \theta_s) & \text{si } E(Q_j, C_s) = 1 \\ 1 & \text{en otro caso} \end{cases} \quad (5.27)$$

Ψ_- es el conjunto de ítems del banco ($\Psi_- \subseteq \Pi$) que no han sido administrados todavía y σ^2 la varianza. $\overrightarrow{u_j}$ es el patrón que podría seleccionar el examinando como respuesta al ítem seleccionado. $P(\theta_s | \overrightarrow{\mathbf{u}_{i-1}})$ es la distribución de su conocimiento a priori sobre el concepto C_s , esto es, antes de que el alumno responda al nuevo ítem seleccionado; y $\rho_w(\theta_s | \overrightarrow{\mathbf{u}_{i-1}}, \overrightarrow{u_j})$ la distribución de su conocimiento a posteriori sobre C_s (después de administrarle el ítem candidato Q_j), asumiendo que el sujeto seleccionará el patrón de respuestas w-ésimo. Por último, nótese que v_{jsw} , en caso de que el ítem Q_j evalúe a C_s , es igual al producto escalar de la distribución de conocimiento a priori por la CCR correspondiente al patrón w-ésimo.

Si se comparan las ecuación 2.32 y la 5.25, puede apreciarse que la primera se restringe únicamente a modelos dicotómicos, en los que las respuestas posibles son 0 (incorrecta) o 1 (correcta). Por el contrario, en esta segunda, se tienen en cuenta las diversas combinaciones de respuesta que podría seleccionar el alumno. Igualmente, también se considera en esta nueva formulación del método de selección, que la respuesta al ítem permite inferir el conocimiento del alumno en más de un concepto.

Es necesario reseñar además que, el modo de evaluación elegido, condicionará el conjunto de conceptos que serán evaluados, y por consiguiente, condicionará el conjunto de ítems que formarán parte del test. Por esta razón, en función de los conceptos que formen parte del conjunto φ , se pueden definir tres modalidades de este criterio:

- a) *Método agregado de selección bayesiana*, en el que hay t conceptos evaluados directamente por el test T_s , esto es:

$$\forall C_j, \quad C_j \in \varphi \quad \Rightarrow \Phi_D(T_s, C_j) = 1 \quad (5.28)$$

- b) *Método completo de selección bayesiana*, en el que hay t conceptos evaluados directamente e indirectamente hacia abajo, esto es:

$$\forall C_j, \quad C_j \in \varphi \quad \Rightarrow \Phi_D(T_s, C_j) = 1 \vee \Phi_{I\downarrow}(T_s, C_j) = 1 \quad (5.29)$$

- c) *Método completo con retropropagación de selección bayesiana*, en el que los t conceptos son todos los evaluados, bien sea directa o indirectamente:

$$\forall C_j, \quad C_j \in \varphi \quad \Rightarrow \Phi(T_s, C_j) = 1 \quad (5.30)$$

Método basado en la dificultad

Este método es una modificación, realizada para este modelo de diagnóstico, del criterio basado en la dificultad de Owen (1969, 1975). La adaptación de este criterio se basa en convertir este método, originalmente aplicado en una única fase, en uno de dos etapas. Consecuentemente, esta nueva versión selecciona primeramente aquel concepto involucrado en

el test, cuya estimación es menos precisa, y posteriormente, elige aquel ítem cuya dificultad está a una distancia menor del nivel de conocimiento estimado de examinando en ese concepto. La precisión de la estimación se evalúa en términos de la varianza de la distribución, ya que cuanto mayor sea ésta, mayor será la dispersión de la distribución. Formalmente, el procedimiento llevado a cabo por este método de selección, se puede expresar de la siguiente forma:

(1) *Selección del concepto C_s :*

$$\max_{C_s \in \varphi} \sigma^2(P(\theta_s | \overrightarrow{\mathbf{u}_{i-1}})) \quad (5.31)$$

(2) *Selección del ítem Q_j :*

$$\min_{Q_j \in \Psi_-} d(b_j, N), \quad \exists C_s, \quad C_s \in \varphi, \quad E(Q_j, C_s) = 1 \quad (5.32)$$

donde

$$N = EAP(P(\theta_s | \overrightarrow{\mathbf{u}_{i-1}}))$$

o bien

$$N = MAP(P(\theta_s | \overrightarrow{\mathbf{u}_{i-1}}))$$

en función del método de inferencia del nivel de conocimiento del alumno utilizado en el test; y

$$d(a, b) = |a - b|$$

La *dificultad de un ítem* (b_j) es uno de los parámetro que caracteriza a las CCI de los modelos logísticos dicotómicos. Aunque son posibles diversas definiciones, aquí se asumirá la proporcionada en la sección 2.8.1, según la cual la dificultad es el nivel de conocimiento para el cual el valor de la probabilidad de responder correctamente al ítem, es el mismo que el de responder de forma incorrecta. Es decir, será aquel nivel de conocimiento cuya probabilidad sea la media de la CCI. Formalmente, esto se puede calcular mediante la siguiente expresión:

$$b_j = \min_{0 < k < K} \left| P_i(R_i(\overrightarrow{u_i}) = 1 | \theta = k) - \frac{P_i(R_i(\overrightarrow{u_i}) = 1 | \theta = K - 1) - P_i(R_i(\overrightarrow{u_i}) = 1 | \theta = 0)}{2} \right| \quad (5.33)$$

Obsérvese que, de forma análoga al anterior, este método de selección tendrá tres modalidades correspondientes dependiendo del criterio de evaluación elegido en el test, y por tanto, en función de los conceptos cuyos ítems intervienen en el test. Esto es:

- a) *Método agregado de selección basado en la dificultad*, si los conceptos del conjunto φ satisfacen la condición expresada en 5.28;
- b) *Método completo de selección basado en la dificultad*, cuando cumplan la condición recogida en 5.29;
- c) *Método completo con retropropagación de selección basado en la dificultad*, en el caso en el que satisfagan la condición expresada en 5.30.

Método basado en la entropía

La *entropía* puede considerarse una medida de la uniformidad de una distribución (Rudner, 2002). Así, cuando ésta es constante, la entropía es máxima. El objetivo en un test adaptativo es tener una distribución de conocimiento del alumno con forma apuntada, es decir, con un máximo claramente diferenciable. Por este motivo, un buen criterio de selección será aquél que seleccione el ítem que contribuya a que el conocimiento del alumno, tras responderlo, tenga la menor entropía. En este método, se aplica esa idea, utilizando el concepto de *entropía esperada*, que se calcula multiplicando el producto escalar de la distribución de conocimiento tras responder a un determinado ítem, por la entropía esperada condicional (en función de la respuesta proporcionada). La formulación de este método de selección es análoga a la utilizada para el criterio de la máxima precisión esperada. Por lo tanto, dado un examinando, que está realizando un test que evalúa a un subconjunto φ de t conceptos $\varphi = \{C_1, C_2, \dots, C_t\}$, con $\varphi \subseteq \Omega$; supóngase que el individuo ha respondido previamente a $i - 1$ ítems; y que $\vec{u}_{i-1} = \{\vec{u}_1, \vec{u}_2, \dots, \vec{u}_{i-1}\}$ es la matriz de los patrones de respuesta que dio a cada ítem administrado. Para calcular cuál es el siguiente ítem Q_j que debe mostrarse en i -ésima posición, por cada ítem del banco (que no haya sido administrado todavía), con W posibles patrones de respuestas $\{\vec{r}_{i0}, \vec{r}_{i1}, \dots, \vec{r}_{iW}\}$, se calcula la entropía esperada de la distribución del conocimiento del alumno, suponiendo que el ítem elegido es Q_j . Se seleccionará aquel con el que se consiga una menor entropía esperada. Formalmente, esto se puede expresar como sigue:

$$\min_{Q_j \in \Psi} \sum_{s=1}^t \sum_{w=0}^W H[\rho_w(\theta_s | \vec{u}_{i-1}, \vec{u}_j)] v_{jsw} \quad (5.34)$$

donde ρ_w y v_{jsw} han sido definidas en las ecuaciones 5.26 y 5.27, respectivamente. Asimismo, la *entropía de una distribución* $H[P(\theta_s)]$ se define de la siguiente forma:

$$H[P(\theta_s)] = \sum_{k=0}^{K-1} -P(\theta_s = k) \log_2 P(\theta_s = k) \quad (5.35)$$

Este criterio se basa en el método de la ganancia de información definido por Rudner *op. cit.* para ítems dicotómicos, en su teoría de los tests adaptativos, dentro de la *Teoría de la Información*, que fue presentado en el capítulo 3. Este método se ha adaptado al modelo de respuesta basado en la TRI propuesto en esta tesis.

Al igual que en los dos criterios anteriores, en función del modo de evaluación del test, se pueden definir tres versiones diferentes:

- a) *Método agregado de selección basado en la entropía*, si los conceptos del conjunto φ satisfacen la condición expresada en 5.28;
- b) *Método completo de selección basado en la entropía*, si cumplen la condición indicada en 5.29;
- c) *Método completo con retropropagación de selección basado en la entropía*, cuando satisfacen la condición expresada en 5.30.

Método basado en la máxima información

Este método se basa en calcular el ítem cuya función de información es máxima para el nivel de conocimiento actual del alumno. Este criterio es el más popular tanto en modelos dicotómicos como en los politómicos (Hontangas et al., 2000). Una de las razones de esta popularidad, es su facilidad de aplicación, puesto que las funciones de información pueden calcularse a priori para todos los ítems. Por este motivo, si antes de comenzar el test ya se tienen calculadas, aplicarlo es sustituir el valor correspondiente al nivel de conocimiento estimado del alumno en cada función de información, y ver para qué ítem el resultado es mayor.

Como se mostró en la sección 2.9.3, existen diversos criterios basados en la información para los modelos politómicos. En el modelo de respuesta presentado en este capítulo, la función de información del ítem se calcula según la propuesta de Dodd et al. (1995), formulada anteriormente en la ecuación 2.31.

Al igual que ocurría con el método basado en la dificultad, este criterio tal cual, no es capaz de hacer una selección de ítems balanceada en contenido (en tests de múltiples conceptos). Por este motivo, debe aplicarse en dos pasos. En el primero de ellos se elige, de entre el conjunto de conceptos φ del test, aquél cuya estimación del conocimiento sea la menos precisa. Una vez seleccionado el concepto C_s , de entre los ítems que lo evalúan, se elige aquél cuyo valor de la función de información para el nivel de conocimiento estimado en el concepto, sea máximo.

El problema está en que, en este modelo, la función de información debe modificarse para tener en cuenta el hecho de que un ítem proporciona evidencias sobre más de un concepto. Por esta razón, la función de información que se ha definido, no sólo tiene en cuenta el ítem sino también el concepto para el que éste se utiliza. Un ítem tendrá, por tanto, tantas funciones de información como conceptos pueda evaluar. Formalmente, el proceso de selección se puede expresar como sigue:

(1) *Selección del concepto C_s :*

$$\max_{C_s \in \varphi} \sigma^2(P(\theta_s | \overrightarrow{\mathbf{u}_{i-1}})) \quad (5.36)$$

(2) *Selección del ítem Q_j :*

$$\max_{Q_j \in \Psi_-} I_{js}(\theta_s) \quad (5.37)$$

donde θ_s es el valor, según la estimación actual del nivel de conocimiento del examinando en el concepto C_s ; y donde la función de información $I_{js}(\theta_s)$ del ítem j para C_s se calcula de la siguiente forma:

$$I_{js}(\theta_s) = \begin{cases} \sum_{w=0}^W \frac{P'_{j\bar{u}_w}(\bar{u}_w | \theta_s)^2}{P_{j\bar{u}_w}(\bar{u}_w | \theta_s)} & \text{si } E(Q_j, C_s) = 1 \\ 0 & \text{en otro caso} \end{cases} \quad (5.38)$$

siendo $P'_{j\bar{u}_w}(\bar{u}_w | \theta_s)$ la derivada de la CCR del patrón de respuesta \bar{u}_w para Q_j y C_s .

Análogamente al resto de métodos de selección de ítems, en función del modo de evaluación del test, se pueden definir las tres versiones diferentes:

- a) *Método agregado de selección basado en la máxima información*, si los conceptos del conjunto φ satisfacen la condición expresada en 5.28;
- b) *Método completo de selección basado en la máxima información*, cuando cumplen la condición indicada en 5.29;
- c) *Método completo con retropropagación de selección en la máxima información*, para el caso en el que satisfacen la condición señalada en 5.30.

De los cuatro criterios de selección anteriores, el primero de ellos, el método bayesiano de la máxima precisión esperada, elige como siguiente ítem, aquél que minimiza la esperanza de la distribución de conocimiento del alumno a posteriori, es decir, después de que lo haya respondido. El objetivo es, por tanto, elegir el ítem que contribuya en mayor medida a minimizar la varianza de su distribución de conocimiento. Cuanto mayor es ésta, mayor es la incertidumbre, ya que la distribución es más dispersa, con lo que ninguno de los niveles tiene una probabilidad lo suficientemente diferenciada del resto. De esta forma, lo ideal es que con cada ítem administrado, la distribución del conocimiento vaya teniendo menor varianza. Esto significará que las probabilidades de un cierto subconjunto de niveles se irán incrementando con respecto al resto. Esta diferenciación afectará cada vez a menos niveles hasta que uno de ellos tenga una probabilidad notablemente mayor que el resto. Como efecto colateral, los criterios de finalización del test estarán más cercanos de satisfacerse, con lo que se asegura la convergencia del algoritmo.

El segundo criterio es similar al anterior. Aunque más eficiente desde el punto de vista computacional, es algo más impreciso. Se ha denominado criterio basado en la dificultad más cercana y, como su nombre indica, selecciona aquel ítem cuya dificultad está a menor distancia del nivel de conocimiento estimado del alumno. Owen (*op. cit.*) demostró que, bajo determinadas hipótesis (véase la sección 2.9.1), este método es equivalente al anterior.

El tercero, el basado en la entropía, es bastante similar al primero. La diferencia radica en la medida de la dispersión de la distribución de conocimiento. Mientras que en el primero se mide en términos de la varianza, en el segundo se hace en función de la entropía de la distribución.

Por último, el cuarto criterio se basa en calcular el ítem que es más informativo para el alumno en ese momento, midiéndose utilizando su función de información en el punto correspondiente a la estimación actual del nivel de conocimiento del examinando.

5.3.3. Criterios de finalización del test

Determinan cuándo debe concluirse la presentación de ítems al alumno en el test. Como se ha mostrado en capítulos anteriores, el criterio de finalización ideal en tests adaptativos es aquél que asegura una evaluación precisa, requiriendo para ello el menor número de ítems.

En este modelo, se han definido tres criterios adaptativos. En todos ellos se analizan las distribuciones del conocimiento del alumno en los conceptos evaluados en el test, después de que éste haya respondido a un ítem. El objetivo es determinar si se cumplen las condiciones requeridas para su terminación. Cada criterio tendrá tres versiones dependiendo del modo de evaluación utilizado en el test, es decir, en función de los conceptos cuyo conocimiento está siendo diagnosticado.

- *Criterio basado en la probabilidad mínima*: Establece que el test debe finalizar cuando la probabilidad asociada al nivel de conocimiento estimado del alumno supere un

umbral. Formalmente, sea un test de evaluación T_s de una asignatura ($T_s \in \Theta$), que evalúa a un conjunto φ de t conceptos $\varphi = \{C_1, C_2, \dots, C_t\}$, sea cual fuere el modo de evaluación. Sea también δ un umbral, y $\vec{\mathbf{u}_i}$ el vector de patrones de respuesta dados a los ítems que han sido administrados hasta el momento. La condición que debe satisfacerse para que finalice el test, se puede expresar de la siguiente forma:

$$\forall C_j, \quad C_j \in \varphi, \quad P(\theta_j = MAP(P(\theta_j | \vec{\mathbf{u}_i})) | \vec{\mathbf{u}_i}) > \delta \quad (5.39)$$

donde el MAP se calcula según la ecuación 5.24. Obsérvese que este criterio se satisface cuando en todas las distribuciones de conocimiento de los conceptos evaluados en el test, su probabilidad máxima está por encima del umbral δ .

- *Criterio basado en la precisión de la estimación:* El objetivo es obtener una distribución de conocimiento con varianza mínima. Si se observa el significado de la varianza de una distribución, analíticamente se puede apreciar que cuanto menor es la varianza de una distribución, más apuntada es la forma de la curva que la describe. Esto indica que la probabilidad del nivel de conocimiento estimado del alumno es mayor en distribuciones con varianzas pequeñas que en aquéllas con varianzas grandes. Así, si la varianza de la distribución de conocimiento del alumno es menor que el umbral (en este caso de varianza mínima), se satisface el criterio de finalización.

Este método puede expresarse formalmente de la siguiente manera: Sea un test de evaluación T_s sobre el conjunto de conceptos φ , δ el umbral del test para este criterio, y $\vec{\mathbf{u}_i}$ el conjunto de patrones de respuesta elegidos por el examinando, la condición de satisfacción del criterio de varianza mínima sería la siguiente:

$$\forall C_j, \quad C_j \in \varphi, \quad \sigma^2(P(\theta_j | \vec{\mathbf{u}_i})) < \delta \quad (5.40)$$

- *Criterio basado en la máxima precisión alcanzable:* Intenta evitar situaciones en las que el nivel de conocimiento del alumno se encuentra entre dos valores contiguos k_i y k_{i+1} . Como consecuencia, en estas situaciones, el modelo no es capaz de discernir si este nivel es k_i , o si por el contrario es k_{i+1} ; y la distribución del conocimiento del examinando variará continuamente, de forma que unas veces el nivel estimado será k_i , y otras veces k_{i+1} , pero en ninguno de los dos casos la precisión de estas estimaciones será suficiente para que se satisfaga la condición de terminación. Así, mediante este criterio, lo que se hace es establecer un intervalo de confianza para la precisión de las estimaciones (δ, ω), de manera que si éstas se encuentran en ese intervalo, se considerarán adecuadas.

Este criterio es similar al anterior. La única diferencia estriba en que, en vez de definir un único umbral δ de precisión, se definen dos: δ y ω . Asimismo, debe cumplirse que ω sea estrictamente mayor que δ , esto es: $\omega > \delta$. De esta forma, una vez que la varianza de la distribución de conocimiento del alumno en un concepto C_t sea menor que ω , y mientras que el valor de esa varianza siga una tendencia decreciente, el test continúa hasta que ésta sea menor que δ . Si por el contrario, en algún momento vuelve a crecer, automáticamente el test finaliza. Este método, al igual que los anteriores, se aplica para todas las distribuciones de conocimiento del alumno en los conceptos evaluados en el test. Su condición de satisfacción puede formularse como sigue:

$$\forall C_j, \quad C_j \in \varphi, \quad \sigma^2(P(\theta_j | \vec{\mathbf{u}_i})) < \delta \quad \vee \quad [\sigma^2(P(\theta_j | \vec{\mathbf{u}_i})) > \sigma^2(P(\theta_j | \vec{\mathbf{u}_{i-1}})) \quad \wedge \quad \delta < \sigma^2(P(\theta_j | \vec{\mathbf{u}_i})) < \omega] \quad (5.41)$$

donde $P(\theta_j|\vec{\mathbf{u}}_{i-1})$ es la distribución de conocimiento del examinando en el concepto C_j tras administrarle $i - 1$ ítems, y $P(\theta_j|\vec{\mathbf{u}}_i)$ su distribución en el mismo concepto tras administrarle i ítems.

Para los tres criterios, el profesor que construye el test debe estimar los valores asociados con los umbrales de decisión. En el primero de los métodos, el valor del umbral deberá ser cercano a uno, mientras que en el segundo y tercero, deberán ser valores cercanos a cero.

Es necesario indicar que, en general, estos criterios adaptativos son convergentes, asegurando por tanto la finalización del test, siempre que se asuman bancos de ítems correctamente construidos. En cualquier caso, es recomendable complementar el criterio de finalización adaptativo elegido con uno adicional (no adaptativo), que evite además una sobreexposición de los ítems, o que la duración del test se prolongue demasiado. Habitualmente se suelen aplicar los dos siguientes criterios no adaptativos:

- *Criterio del máximo número de ítems:* Según el cual, el test finaliza cuando el número de ítems administrado sobrepasa un determinado umbral, independientemente de si se cumple o no el criterio de finalización adaptativo.
- *Criterio temporal:* Se basa en el establecimiento a priori de un tiempo de duración del test. El criterio de finalización sería, o bien un criterio adaptativo, o bien haber alcanzado el tiempo máximo permitido.

Desde los principios de la psicometría, ha habido diversos intentos de utilizar el tiempo empleado por el alumno en responder a un ítem, como un componente adicional de la propia respuesta. Estudios como el realizado por Hornke (2000) avalan la hipótesis de que el tiempo de respuesta es mayor cuando la respuesta es errónea. Así, existen modelos que utilizan este valor como parte de la corrección del ítem (White, 1982), y más recientemente otros que restringen el tiempo total de respuesta (Roskam, 1997; Verhelst et al., 1997). A pesar de ello, actualmente, en el modelo de respuesta propuesto en esta tesis no se tiene en cuenta este aspecto, únicamente se puede utilizar como mecanismo de control de la finalización del test.

5.4. Calibración de las curvas características del modelo

En el capítulo anterior se explicó cómo se lleva a cabo la calibración de los ítems según el modelo de respuesta propuesto. El modelo de diagnóstico presentado en este capítulo introduce nuevas relaciones entre un ítem y el conjunto de conceptos a los que evalúa. Como consecuencia de ello, es necesario modificar sensiblemente el procedimiento de calibración para adaptarlo a esta nueva situación. Ahora, cada combinación ítem/opción de respuesta tiene tantas CCO como conceptos evalúa el ítem, tal y como refleja la figura 5.3. Por consiguiente, el procedimiento de calibración ya no sólo se aplica una vez por test, sino que deberá repetirse por cada concepto evaluado en el mismo. De esta forma, cada test puede considerarse, desde el punto de vista de la calibración, como una colección de t subtests, donde t es el número de conceptos evaluados. El algoritmo de calibración se aplicará, siguiendo el procedimiento explicado en el capítulo anterior, con la única variación de que se hace un número de veces igual al producto del número de ítems que se desea calibrar, multiplicado por el número de conceptos que éstos evalúan, bien sea directa o indirectamente.

Sea Φ_c el subconjunto de ítems que se desean calibrar, tal que: $\Phi_c \subseteq \Phi$. A partir de él, se determina el conjunto de tests Π_c , tal que: $\Pi_c \subseteq \Pi$. Π_c está formado por todos aquellos tests en los que se han administrado los ítems de Φ_c . Una vez determinado Π_c , se averigua el conjunto de conceptos que han sido evaluados en Π_c . Llámese a ese conjunto de conceptos Ω_c , tal que: $\Omega_c \subseteq \Omega$.

Por cada concepto C_j incluido en Ω_c ($C_j \in \Omega_c$), se toma el subconjunto de tests Π_{cj} , $\Pi_{cj} \subseteq \Pi_c$, tal que sus elementos deben cumplir la siguiente condición:

$$\forall T_s, \quad T_s \in \Pi_{cj}, \quad E_D(T_s, C_j) + E_{I\downarrow}(T_s, C_j) = 1 \quad (5.42)$$

Es decir, se toman aquellos tests que evalúan directa o indirectamente hacia abajo al concepto C_j . Asimismo, se extraen las sesiones realizadas de esos tests, y se calcula la calificación (heurística) obtenida por cada examinando en C_j , en el test correspondiente. Este cálculo anterior, supone aplicar la primera fase del algoritmo de calibración basado en el suavizado núcleo. A partir de esas evaluaciones, se prosigue con el algoritmo, tal y como se describió en la sección 4.7. Una vez finalizada la calibración, como resultado se obtendrán las CCO de los ítems para el concepto C_j , así como las estimaciones de los niveles de conocimiento sobre C_j de los alumnos de la muestra.

El algoritmo se deberá aplicar, de forma análoga, para todos los conceptos de Ω_c . Esto implica, por tanto, que se deberá repetir un número de veces igual al producto de los cardinal de Φ_c y de Ω_c , esto es: $\zeta(\Phi_c) \times \zeta(\Omega_c)$.

5.5. Conclusiones

En este capítulo se ha presentado un modelo de evaluación cognitiva basado en TAI para el diagnóstico el STI (Guzmán y Conejo, 2004a). Esta propuesta se cimenta sobre el modelo de respuesta basado en la TRI introducido en el capítulo anterior. Su presentación se ha realizado en dos partes: En la primera de ellas se ha mostrado la arquitectura del modelo, que combina los elementos necesarios para administrar TAI, con los necesarios para el diagnóstico en STI. En el modelo, el dominio sobre el que el alumno es evaluado se basa en una red semántica de conceptos.

Se ha definido de forma genérica la noción de concepto, como todo aquello sobre lo que se puede evaluar al alumno. Los conceptos se relacionan entre sí mediante relaciones de agregación. La red semántica de conceptos, junto con los ítems y las especificaciones de tests forman el módulo experto del modelo de diagnóstico.

En la segunda parte se ha descrito el funcionamiento del modelo, que se basa en la administración de TAI multiconceptuales, en los que se aplican criterios de selección de ítems y de finalización adaptados a las características del modelo. Se han propuesto cuatro criterios adaptativos de selección de ítems, que tienen en cuenta las características polítómicas y multiconceptuales del modelo de diagnóstico. Esto implica que estos métodos, sin necesidad de recurrir a heurísticos, son capaces de realizar una selección de ítems balanceada en contenido. Asimismo, se han definido tres criterios adaptativos de finalización, que se emplean para determinar si el modelo tiene las evidencias suficientes para diagnosticar, de forma adecuada, el conocimiento del alumno en los conceptos que intervienen en el test.

Con respecto a los modelos basado en los TAI tradicionales, esta aproximación tiene la ventaja de que permite la evaluación simultánea de múltiples conceptos dentro de un

mismo test. Esta característica lo hace idóneo para su aplicación al diagnóstico del alumno. Aunque en principio pudiera parecer que se está violando una de las condiciones que deben satisfacerse cuando se administran TAI, esto es, la condición de unidimensionalidad, realmente no es así. Los tests del modelo en los que se evalúan diversos conceptos directamente, son como una colección de pequeños tests (uno por cada concepto) que se administran de forma simultánea. Asimismo, cuando se realizan tests configurados con el modo de evaluación completa, el problema puede verse como una evaluación multicomponente, y como consecuencia, el modelo puede considerarse como un modelo multicomponente.

A priori, un inconveniente de este modelo es la cantidad de curvas características que utiliza, y que por tanto, deben ser calibradas antes de poder administrar tests adaptativos. Para contrarrestar este problema, se hace uso de un método de calibración cuyos requisitos son menores que los de los métodos tradicionalmente utilizados con este fin en el ámbito de la TRI. Las ventajas del uso de este algoritmo de calibración se evaluarán en el capítulo 7.

En el capítulo siguiente se presenta la implementación que se ha llevado a cabo de este modelo de evaluación cognitiva, dentro de un sistema para la construcción y generación de tests a través de la web, denominado SIETTE.

Parte IV

IMPLEMENTACIÓN

Capítulo 6

Implementación del modelo: El sistema SIETTE

*No basta saber,
se debe también aplicar.
No es suficiente querer,
se debe también hacer.*
Goethe

El modelo de evaluación cognitiva presentado en el capítulo anterior de esta tesis, está actualmente implementado en el sistema SIETTE¹ (*Sistema Inteligente de Evaluación mediante Tests para Teleeducación*), un entorno de aplicaciones web para la construcción y generación de tests, cuya página principal se muestra en la figura 6.1. Principalmente, permite construir y administrar TAI, aunque es un sistema que también ofrece la posibilidad de administrar tests convencionales basados en la TCT (Conejo y Guzmán, 2002).

A través de una interfaz web, los alumnos pueden realizar tests para autoevaluarse (Guzmán y Conejo, en prensa), donde la corrección de cada ítem se muestra inmediatamente después de cada respuesta, y opcionalmente un refuerzo; o bien los profesores pueden evaluar a sus alumnos, utilizando SIETTE como medio de evaluación académica.

Para construir y modificar los tests así como su contenido, SIETTE ofrece a los profesores un entorno de herramientas de autor. Este entorno permite además analizar las sesiones de tests llevadas a cabo por los alumnos, así como la calibración de los ítems.

Este sistema puede funcionar bien como una herramienta de evaluación independiente, o integrado dentro de otros sistemas web educativos, especialmente en sistemas de enseñanza adaptativos, en los que SIETTE puede desempeñar el rol de sistema de diagnóstico del conocimiento del alumno. Es además multilingüe (actualmente disponible en español, inglés y alemán) y abierto a la inclusión de otros idiomas, gracias a la generación dinámica de sus interfaces.

Dentro de este capítulo, en primer lugar, se van a presentar los tipos de ítems que incluye SIETTE, que son a su vez la implementación de los presentados en el modelo de respuesta propuesto en el capítulo 4. A continuación se hace una descripción de la base

¹La dirección a través de la cual se puede acceder a este sistema es: <http://www.lcc.uma.es/SIETTE>.

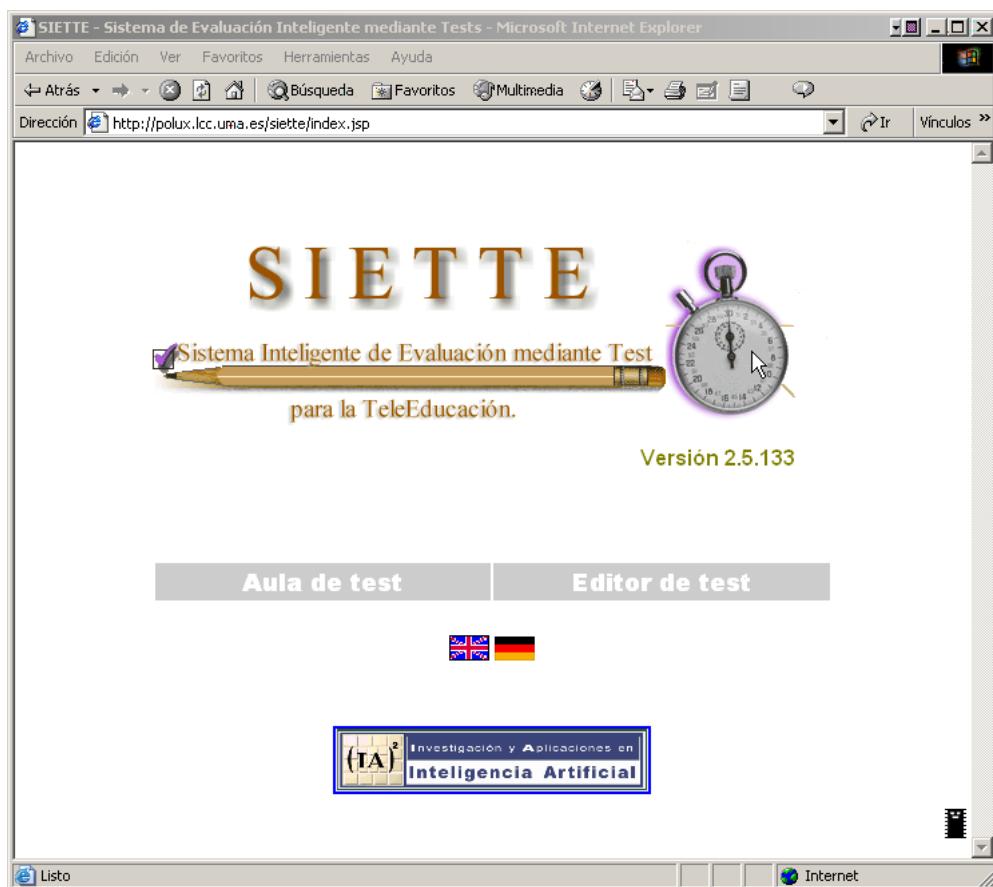


Figura 6.1: Entrada al sitio web de SIETTE.

de conocimientos, que es la implementación del módulo experto del modelo de diagnóstico cognitivo presentado en el capítulo anterior. Posteriormente, se describen los repositorios de los modelos de los alumnos y de los profesores. Tras esto, se presentarán las herramientas que manejan los usuarios finales. Primeramente, la que se ha denominado *aula virtual*, que permite al alumno realizar tests, y la segunda, el *editor de tests*, aplicación que ayuda al profesor a construir sus asignaturas y los tests con los que podrá evaluar a sus alumnos. Se describirán también las herramientas de apoyo al sistema, que son el analizador de resultados y el calibrador de ítems. Asimismo, se abordará el funcionamiento de la interfaz de conexiones externas, que permite integrar SIETTE con otros sistemas, a través de un protocolo de interacción, especialmente diseñado con este objetivo. Finalmente es presentada la arquitectura del sistema, algunos detalles de implementación y una breve descripción de la evolución del sistema.

6.1. Tipos de ítems en SIETTE

SIETTE es un sistema cuya interfaz de desarrollo es la Web. Por este motivo, desde el punto de vista de la implementación, los ítems pueden verse como fragmentos de código HTML. Tanto el enunciado como el texto asociado a cada opción de respuesta son fragmentos de HTML que, en el momento de la presentación del ítem, se componen dinámicamente en la página web encargada de mostrarlo al alumno. Esto amplia enormemente las posibilidades en cuanto a la inserción de contenidos se refiere. Por consiguiente, un enunciado o un texto de una opción de respuesta pueden contener imágenes, sonidos, videos, etc. En resumen, todo aquello que pueda ser incluido dentro de una página web.

SIETTE maneja todos los tipos de ítems que se describen a continuación, junto con su mecanismo de evaluación. Adicionalmente, los ítems podrán ser de *siettlets* (que serán definidos en la sección 6.1.2), generativos, temporizados o externos, de manera que ninguna de estas categorías son excluyentes entre sí. De esta forma, un ítem podrá ser, por ejemplo, de opción múltiple (tipo básico), *siettlet*, generativo y temporizado.

6.1.1. Ítems básicos

Como implementación del modelo de diagnóstico, SIETTE permite a los profesores construir tests con los ítems presentados dentro del modelo de respuesta definido en el capítulo 4. A este conjunto de ítems, junto con los ítems de respuesta corta (que serán presentados en la siguiente sección) se les denomina *ítems básicos (predefinidos o internos)*, puesto que cualquier tipo de ítem adicional que se puede definir en SIETTE, desde el punto de vista del modelo de respuesta, corresponderá a alguno de estos ítems.

SIETTE permite, por tanto incluir: ítems verdadero/falso, de opción múltiple, de respuesta múltiple con opciones dependientes y con opciones independientes, de ordenación y de relación. La figura 6.2 (que será descrita en detalle en la sección 6.5) muestra la apariencia de los ítems en SIETTE durante la fase de administración de un test al alumno. Cualquier tipo de ítem que se incluye en SIETTE, aunque en su apariencia pueda ser diferentes, siempre equivaldrá, desde el punto de vista de la evaluación, a uno de los tipos de ítems definidos en el modelo de respuesta. Los ítems verdadero/falso, de opción múltiple, de respuesta múltiple, de ordenación y de relación, pueden incluirse en el formato convencional similar al que se utiliza en los tests de papel y lápiz, tal y como se puede apreciar en la figura 6.3. Además, se han definido otros tipos de ítems con formatos más innovadores, los cuales se enumeran a continuación en las subsecciones siguientes.

Ítems de respuesta corta

En este tipo de ítems, el alumno debe escribir la respuesta (o respuestas) que satisfacen un determinado enunciado. Las opciones de respuestas se representan mediante patrones. Podrá haber tantos patrones para identificar respuesta correctas, como para identificar respuestas incorrectas. Además, se incluye un patrón por defecto, que permite identificar una respuesta incorrecta que no satisface ninguno de los patrones.

El procedimiento de evaluación se aplica, en función del tipo de ítem de respuesta corta, por cada una de las respuestas escritas por el alumno, de la siguiente forma: Se comprueba qué patrón de respuesta satisface. Es precisamente la opción de respuesta a la que está asociado ese patrón, la que se marcarán como respuestas seleccionadas por el alumno. En caso

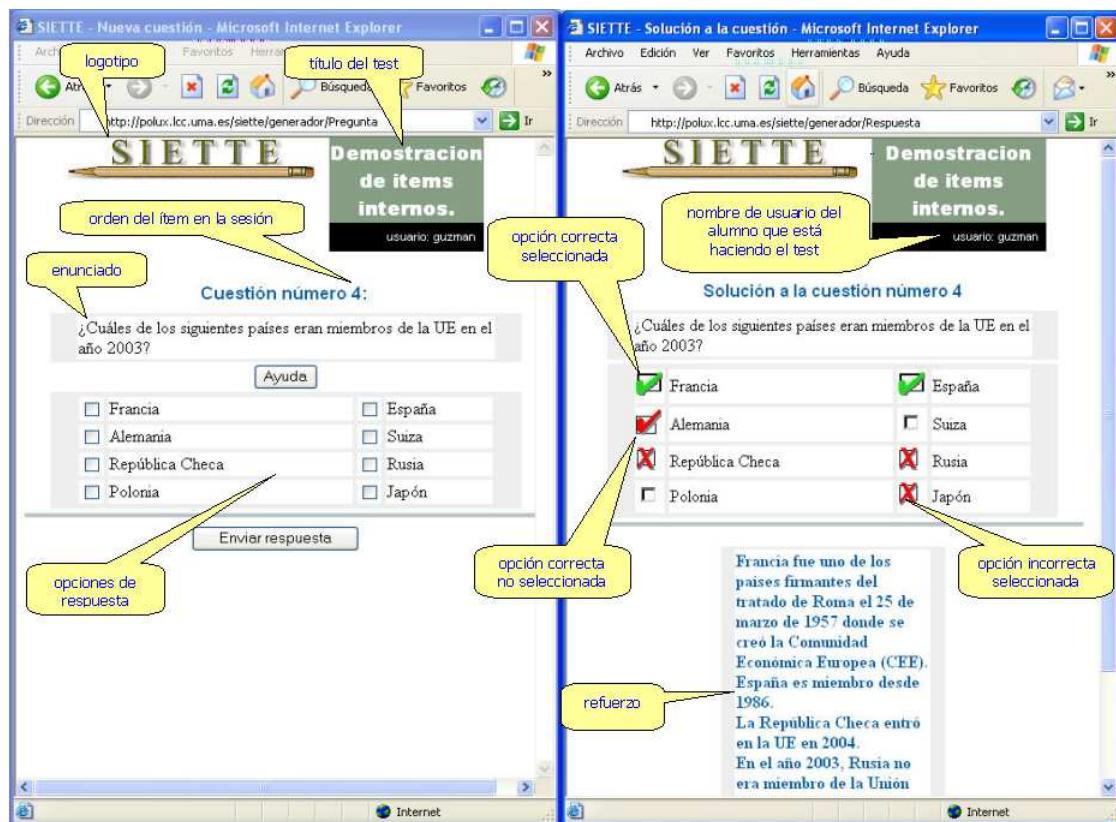


Figura 6.2: Un ítem mostrado en el aula virtual y su corrección.

de que no se satisfaga ninguno de los patrones, se asume que el alumno ha seleccionado la respuesta asociada al patrón de respuesta por defecto.

En función de cómo se corrige la respuesta escrita por el alumno, existen en SIETTE tres tipos de ítems de respuesta corta:

- *Basados en correspondencia*: La respuesta suministrada por el alumno se compara (tal cual) con los patrones de respuesta almacenados.
- *Basados en expresiones regulares*: Durante el proceso de construcción de un ítem de este tipo, se almacenan como respuestas, expresiones regulares. El procedimiento de corrección de un ítem de este tipo consiste en comprobar, por cada patrón de respuesta, si el conjunto de respuestas suministradas por el alumno pertenece al lenguaje generado a partir de la expresión regular correspondiente.
- *Basados en expresiones regulares Java*: Son análogos a los anteriores, aunque ofrecen la posibilidad de incluir expresiones más complejas².

²Se implementan mediante una clase Java diseñada con este fin, que fue incluida a partir de la versión 1.4 de Java.

¿Cuál de los siguientes países está en Europa?

<input type="radio"/> Japón	<input type="radio"/> Congo
<input type="radio"/> USA	<input type="radio"/> Francia

¿Cuáles de los siguientes países eran miembros de la UE en el año 2003?

<input type="checkbox"/> República Checa	<input type="checkbox"/> Polonia
<input type="checkbox"/> España	<input type="checkbox"/> Suiza
<input type="checkbox"/> Alemania	<input type="checkbox"/> Japón
<input type="checkbox"/> Rusia	<input type="checkbox"/> Francia

Figura 6.3: Ítems de opción múltiple y de múltiple respuesta.

A través de este tipo de ítems se pueden almacenar patrones de respuesta que permiten identificar respuestas incorrectas. La principal utilidad de estos patrones incorrectos es, en tests de autoevaluación, identificar los errores más comunes que cometan los alumnos (denominados en inglés *misperceptions*) y generar refuerzos en consecuencia. Durante el proceso de calibración de este tipo de ítems, a partir de las sesiones con las que se lleve a cabo la calibración, se pueden identificar los patrones de respuesta errónea más frecuentes, aplicando un proceso de *bootstrapping*, e incluirlos como patrones de respuesta incorrectos, asociándoles a su vez un refuerzo.

En tests en los que la corrección del ítem se muestra inmediatamente después de que el alumno responda, para el caso en el que el alumno responda incorrectamente, en la corrección se muestra, como respuesta correcta, un ejemplo que debe proporcionar el profesor durante la fase de construcción del ítem.

Cita algún país firmante del tratado de Roma de 1957

Cita tres países diferentes de entre los firmantes del Tratado de Roma de 1957.

Ayuda

Figura 6.4: Ítems de respuesta corta.

Desde el punto de vista del modelo de respuesta y, por tanto, de la evaluación basada en la TRI, este tipo de ítems puede verse como una colección de ítems de opción múltiple (tantos como casillas de respuesta se incluyan en el ítem). En la figura 6.4 se han representado dos ítems de este tipo. En el ítem de la derecha el alumno tendrá que escribir N respuestas, en vez de una sola como en el de la izquierda. Cada una de esas respuestas se evalúa independientemente (como un ítem de opción múltiple) para ver si es correcta, y además se

comprueba que no coincida con ninguna de las otras respuestas introducidas por el alumno para ese mismo ítem.

Si bien es cierto que otros sistemas (especialmente los sistemas comerciales) permiten la posibilidad de incluir este tipo de ítems, tal y como se puso de manifiesto en la sección 2.10; en principio parece que sólo SIETTE permite que la corrección de éstos se haga a partir de expresiones regulares incluidas. En el resto de sistemas, se suelen implementar mecanismos de corrección basados solamente en correspondencia.

6.1.2. *Siettlets* o ítems autocorregidos

El hecho de que los ítems sean fragmentos de HTML permiten también la inclusión en ellos de pequeños programas como applets de Java o archivos Flash embebidos en el código. En virtud de esto, SIETTE permite la inclusión en sus tests de los denominados *siettlets* o *ítems autocorregidos*, que son especialmente útiles cuando la respuesta requiere un nivel alto de interactividad. Estos ítems no ofrecen al alumno un conjunto de opciones de respuestas de entre las que deban elegir aquéllas que consideren correctas. En estos ítems, el alumno debe interactuar con un pequeño programa. En función de la interacción, será el propio programa el que evalúe al alumno y devuelva esta evaluación en forma de opción (u opciones) de respuesta. Por lo tanto, realmente este tipo de ítems se componen de un enunciado, en el que se incluye el pequeño programa, y un conjunto de opciones de respuesta, ocultas al alumno, pero conocidas de antemano por el programa y por SIETTE. Una vez que el alumno termina su interacción con el ítem, éste, automáticamente y de forma completamente transparente, evalúa la actuación del alumno, y selecciona la opción u opciones de respuesta adecuadas, las cuales envía a SIETTE. Por consiguiente, en este tipo de ítems, el procedimiento de corrección habitual se efectúa internamente en *siettlet*, tal y como se explicará posteriormente.

El primer test que se construyó con *siettlets* es el test de Piaget. Este test supuso la primera integración de un sistema externo en SIETTE. El objetivo de este test fue estimar el nivel de desarrollo cognitivo en niños (Arroyo et al., 2001) de entre 8 y 11 años. Se trata de un test adaptativo compuesto por 10 ítems, cada uno de los cuales evalúa una de las siguientes habilidades cognitivas del alumno: serialización, conservación de números, reciprocidad, conservación de áreas, inclusión en clases, funcionalidad, reversibilidad, establecimiento de hipótesis, proporcionalidad y análisis combinatorio. Antes de ser integrados en SIETTE, estos ítems fueron calibrados a partir de sesiones de evaluación realizadas a través del sistema *Animal Watch* (Arroyo et al., 1999, 2000). Posteriormente, esos ítems fueron adaptados para su posterior integración en SIETTE. Atendiendo al tipo de respuesta, todos estos ítems se implementaron como ítems verdadero/falso. Consecuentemente, todos estos ítems tienen únicamente dos respuesta internas: "correcto" e "incorrecto". Si el alumno resuelve el ejercicio correctamente, el applet internamente seleccionará la opción de respuesta "correcto", y en otro caso, la opción de respuesta "incorrecto".

Otro tipo de ítems que se ha construido utilizando *siettlets* son los *de respuesta sobre figura*. Un buen ejemplo de un ítem de este tipo se ha incluido como parte del banco de ítems de un test para la identificación de especies europeas de árboles y su distribución geográfica. La figura 6.5 muestra este ítem que se ha implementado mediante un applet de Java. Su objetivo es, sobre un mapa geográfico de Europa o España, y utilizando el ratón como pincel, seleccionar las regiones en las que pueden encontrarse ejemplares de una determinada especie forestal indicada en el enunciado del ítem. Atendiendo al tipo de respuesta, este ítem es de opción múltiple. Sus opciones de respuestas son los intervalos

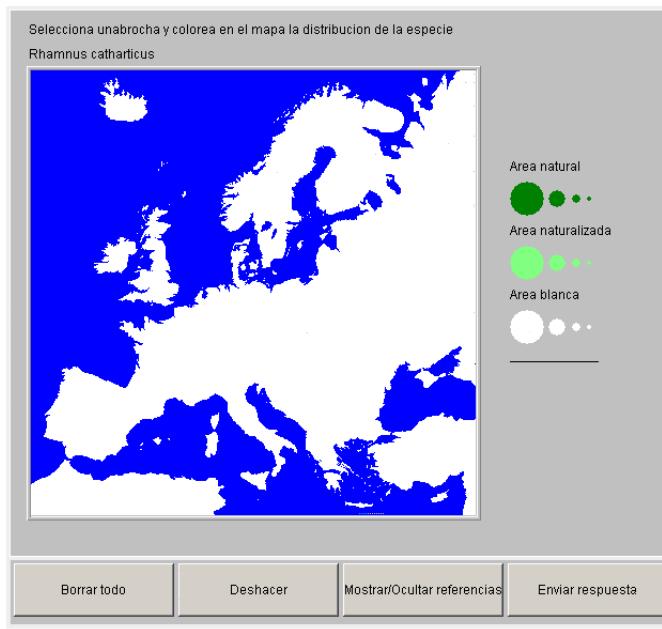


Figura 6.5: Ítem de respuesta sobre figura.

de porcentajes de corrección. Se han incluido cuatro opciones de respuesta: de entre 0 % y 25 %, entre 25 % y 50 %, entre 50 % y 75 %, y por último entre 75 % y 100 %. La opción de respuesta correcta es claramente la última. Una vez que el alumno ha hecho la selección de las regiones con el pincel, deberá pulsar el botón "Corregir". Tras esto, el applet internamente comparará la región seleccionada con la región correcta y determinará el porcentaje de acierto. Posteriormente, el propio applet, en función de ese porcentaje, devolverá a SIETTE la respuesta correspondiente al intervalo en el que se encuentra dicho porcentaje.

La figura 6.6 muestra de forma esquemática, cómo se lleva a cabo el proceso de corrección en este tipo de ítems. Inicialmente se muestra el testlet al alumno. Como se puede apreciar, la lista de opciones de respuesta está oculta. Una vez que haya interactuado con el programa (paso 1 de la figura), éste calculará, a partir de la interacción con el alumno, la correspondiente opción de respuesta (paso 2), y enviará ésta al módulo de evaluación (paso 3), siguiendo el procedimiento habitual. En (Guzmán y Conejo, 2004c) se describe con mayor detalle este mecanismo. Como se ha mencionado, un *siettlet* podrá ser simultáneamente de cualquiera de los tipos de ítems incluidos en el modelo de respuesta.

Gracias a esta técnica, SIETTE ofrece la posibilidad de incluir prácticamente cualquier tipo de ítems (siempre que éste pueda ser implementado mediante un programa que pueda ser embebido en una página web). La inclusión de estos applets no supone ninguna alteración en cuanto al mecanismo de evaluación adaptativa se refiere. El proceso de evaluación es el mismo que se sigue para el resto de ítems. Esto se puede realizar gracias a la comunicación entre el código Javascript de la página de presentación del ítem y el programa. Por este motivo, es también requisito indispensable que el lenguaje de programación en el que esté escrito el programa, no sólo permita incluirlo en una página web, sino también que permita su interacción con código Javascript inmerso en etiquetas HTML.

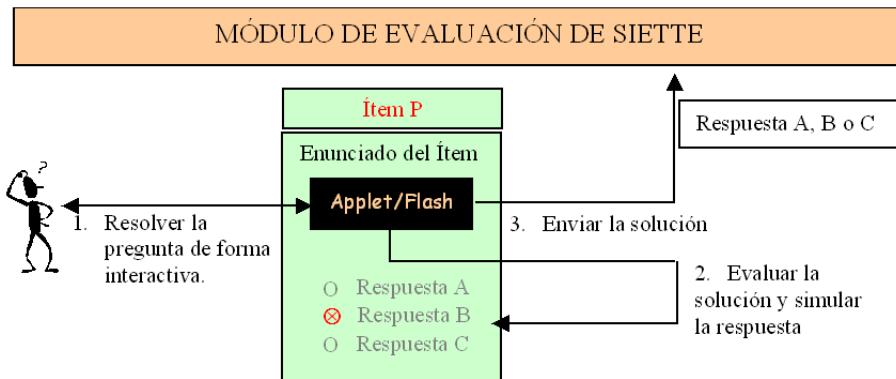


Figura 6.6: Proceso de evaluación seguido por un *siettle*.

6.1.3. Los ítems generativos

Tal y como se ha puesto de manifiesto en capítulos anteriores, uno de los principales inconvenientes en la construcción de un TAI es el tamaño del banco de ítems. Lo deseable es que éste sea lo mayor posible. Flaugh (1990) recomienda que el banco tenga al menos 500 ítems. La razón principal para exigir un banco de ítems de tamaño considerable es evitar, entre otras cosas, la sobreexposición de los ítems. Si se da este caso, puede causar que los alumnos que realicen un test se aprendan sus ítems, y puedan compartir este conocimiento con alumnos que vayan a realizar el test con posterioridad. Para garantizar una evaluación cognitiva fiable, es necesario evitar esto. El problema radica en que ésta es una tarea que requiere mucho tiempo. Una solución parcial a este problema reside en el propio modelo de respuesta que, al ser polítómico implica que el número de ítems de un tests adaptativo será sustancialmente inferior al requerido en TAI sobre modelos de respuesta dicotómicos. Otra solución adicional es la *generación automática de ítems* o también denominada (en el campo de la psicometría) *modelado de ítems* (en inglés, *item modeling*) (Béjar et al., 2003). La idea principal que subyace en la generación automática de ítems es que los profesores, en vez de construir aquéllos directamente, construyan plantillas de ítems, las cuales se instanciarán, generando cada vez, como resultado, un ítem isomorfo diferente. Según Béjar et al. (2003) éstos son *"ítems que tienen un contenido comparable, y cuyas propiedades psicométricas son intercambiables"*.

En (Belmonte et al., 2002) se llevó a cabo un estudio exhaustivo de las técnicas de generación automática de ítems en sistemas sobre la web. La clasificación que se realizó los divide en tres grupos, en función de las técnicas de generación que utilizan: (1) *Basados en mecanismos simples*: Suponen la construcción de plantillas de ejercicios, a partir de lenguajes de programación, que incluyan sentencias de generación pseudoaleatoria de números. (2) *Basados en Gramáticas de Contexto Libre*: Éstas permiten, a partir de un pequeño conjunto de reglas de producción, generar sentencias complejas. (3) *Basados en técnicas ad-hoc*.

El *Sistema de Evaluación Inteligente Bizantino* (Patel et al., 1998) permite la generación de ítems. Este sistema evalúa a alumnos sobre los siguientes conceptos de Economía: la tasa de inversión capital, coste de absorción, coste marginal y coste estándar. Sus autores lo presentan como una arquitectura para la evaluación que se incluye dentro de un conjunto de herramientas de tutorización inteligentes bizantinas. Este sistema se basa en el uso de

cuestiones generativas en las que el alumno debe insertar un conjunto de variables dependientes a partir de un conjunto de variables independientes suministradas en el enunciado. Otro tipo de cuestiones que ofrece este sistema, son de respuesta corta, esto es, ejercicios en los que el alumno debe ser capaz de calcular las variables dependientes para un determinado problema. El mecanismo de evaluación parece estar basado en criterios convencionales que hacen uso de heurísticos que permiten penalización y puntuación parcial de las cuestiones.

Por otro lado, Mavrikis y González Palomo (2003) utilizan una aproximación similar a la utilizada en SIETTE. Mediante un lenguaje de descripción basado en XML especifican ítems de matemáticas, donde se pueden definir variables cuyos valores se calculan dinámicamente y de forma aleatoria. Éstas se introducen en el enunciado del ítem o en las respuestas, dando lugar a diferentes ítems isomorfos cada vez que se generan.

En la herramienta QUIZPACK (*Quizzes for Parameterized Assessment of C Knowledge*), Pathak y Brusilovsky (2002) han desarrollado un sistema de generación de ítems dentro de tests estáticos de entre 5 y 10 cuestiones. El objetivo de estos tests es evaluar el conocimiento de un alumno sobre el lenguaje C/C++. En este sistema, el mecanismo de generación de ítems, aunque similar al que utiliza SIETTE, es bastante más estricto en cuando a desarrollo, ya que el profesor debe compilar las plantillas antes de poder ser utilizadas. Asimismo, cada cambio en una plantilla obliga a repetir el proceso de recompilación.

SIETTE permite la creación de plantillas para la generación de ítems isomorfos en tiempo real. Cuando un profesor utiliza una plantilla en un test, ésta es tratada como un ítem más desde el punto de vista psicométrico. De esta forma, durante la administración de un test, cada vez que el sistema selecciona una plantilla, se genera un ítem isomorfo diferente. La única diferencia con respecto al resto, es que, estos ítems, una vez seleccionados, no son retirados del banco de ítems. Las plantillas se implementan mediante lenguajes embebidos en HTML, tales como JSP o PHP. La inclusión de un ítem generativo es por tanto, realizada de forma completamente transparente. Cuando un profesor quiere insertar un ítem generativo en SIETTE, inicialmente sólo deberá seleccionar el tipo de ítem, como si se tratara de un ítem no generativo. Posteriormente, durante la edición del ítem, deberá indicar que se trata de un ítem generativo, marcando la opción adecuada a través de la herramienta de autor. Adicionalmente, deberá indicar el lenguaje en el que se va a construir la plantilla, así como el número máximo de instancias de la plantilla que pueden aparecer en un mismo test. Como desde el punto de vista del formato, las plantillas también representan un ítem más, cada plantilla tendrá un enunciado y un conjunto de posibles opciones de respuesta. El profesor deberá añadir en el enunciado y en el texto de las opciones de respuesta, sentencias de código en el lenguaje embebido seleccionado. Estas sentencias deberán contener funciones de generación aleatoria de valores, especialmente en el texto de las respuestas, que permitirán que, cada vez que la plantilla sea instanciada, se generen valores diferentes. Como resultado, estas sentencias serán las encargadas de generar las respuestas correctas e incorrectas en función de un determinado enunciado.

Cuando un alumno está realizando un test y el sistema de generación determina que el siguiente ítem que debe mostrarse es de tipo generativo, se lleva a cabo un procesamiento en cuatro fases. El proceso implica, por tanto, la inclusión de tres pasos adicionales, que se añaden al principio del mecanismo habitual de visualización, seguido en el caso de un ítem no generativo:

1. *Empaquetado*: Se construye una página web temporal formada por el enunciado y el texto de las opciones de respuesta.

2. *Preprocesado*: Internamente, esta página es enviada a un servidor web con la tecnología necesaria para poder interpretar el lenguaje en el que está escrita la plantilla. El servidor devuelve a SIETTE la página con el código HTML generado como resultado.
3. *Desempaquetado*: La página devuelta por el servidor web es desensamblada y separada en enunciado y opciones de respuesta.
4. *Visualización*: Al igual que sucede con el resto de ítems, el ítem es mostrado al alumno.

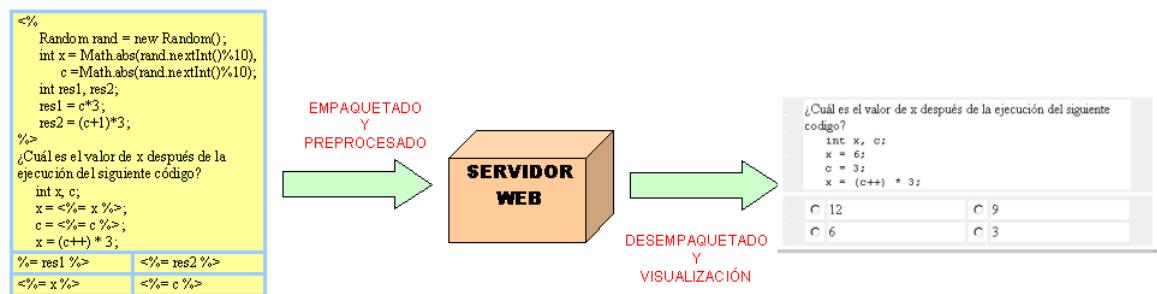


Figura 6.7: Generación y administración de un ítem generado a partir de una plantilla.

El ítem de respuesta sobre figura (mostrado en la figura 6.5) es un ejemplo de este caso. Es una plantilla, en la que el mapa que se muestra al alumno (de España, Europa o Andalucía) se elige de forma aleatoria, cada vez que el ítem es seleccionado. Asimismo, la especie vegetal que el alumno debe localizar en el mapa, también se genera de forma dinámica.

La figura 6.7 muestra otro ejemplo de un ítem generativo en SIETTE, cuya plantilla está escrita en JSP. Se trata de una cuestión sobre el lenguaje de programación Java. El código principal de la plantilla, así como las sentencias de generación aleatoria de valores, se almacenan en el enunciado del ítem. El código que genera cada opción de respuesta se almacena en el texto correspondiente a cada respuesta. Los valores de las variables *x* y *c* son generados cada vez que el ítem generativo es seleccionado. Nótese que este mecanismo permite la generación de distractores cercanos a la respuesta correcta, es decir, respuestas que no siendo correctas son muy similares a la respuesta correcta. En el ítem generado a la derecha de la figura 6.7, la respuesta correcta es 9, pero si el alumno tiene ciertos conocimientos de Java, pero no los suficientes para responder de forma correcta al ítem, existe una probabilidad considerable de seleccionar el distractor 12.

Para ayudar a los profesores en la construcción de este tipo de ítems, la herramienta de autor (sección 6.6) incluye una utilidad que permite forzar la generación de un ítem a partir de una plantilla, y así que el profesor pueda previsualizarlo. Esto facilita considerablemente la construcción de plantillas y la depuración de los posibles errores que puedan surgir en este proceso. El principal problema que supone este mecanismo de generación de ítems isomorfos es la misma que presentan los *siettlets*, esto es, que los profesores deben tener ciertos conocimientos de programación para poder construir correctamente las plantillas.

El uso de ítems generativos podría plantear a priori un problema en cuanto a la calibración. Téngase en cuenta que su aspecto final difiere por cada instancia que del ítem se genera. En este sentido, Béjar et al. (2003) realizaron un estudio sobre la calibración de estos ítems. Este estudio se componía a su vez de dos pruebas con ítems generativos

construidos a partir de otros no generativos calibrados con anterioridad: una de ellas realizadas con alumnos virtuales mediante un simulador, y la segunda de ellas con alumnos reales, utilizando una herramienta para la generación de tests a través de Internet. Después de recopilar la información de las sesiones de tests realizadas por los alumnos reales por una parte, y por los alumnos simulados por otra, se llevó a cabo la calibración de los ítems. El objetivo del estudio era validar la nueva calibración. Para la calibración a posteriori se utilizó el método basado en la *Función de Respuesta Esperada* (en inglés, *Expected Response Function*) (Mislevy et al., 1994), que es una técnica en la que se lleva a cabo una atenuación de los parámetros estimados. Según los autores del estudio, este método es perfecto para este problema, ya que asume que las propiedades psicométricas de un determinado ítem pueden presentar cierta variabilidad cada vez que éste es administrado. Los resultados de este estudio mostraron que no existían diferencias significativas (en cuanto a propiedades psicométricas) entre las calibraciones realizadas a priori y aquéllas realizadas a posteriori. Esto permite garantizar que la inclusión en TAI de ítems generativos no supone ninguna merma considerable en cuanto a la precisión de las estimaciones obtenidas. En la actualidad, SIETTE calibra los ítems generativos como si se tratara de ítems no generativos, no aplicando por tanto, ninguna técnica específica para este tipo de ítems.

6.1.4. La biblioteca de plantillas

Una de las desventajas de los sistemas tradicionales para la generación de tests es que sus ítems tienden a tener siempre el mismo formato: un enunciado y un conjunto de opciones de respuesta. Según Thissen (1993) el principal motivo de esto es que estos sistemas representan una evolución natural de los sistemas de tests de lápiz y papel, basados en la TCT, en los que era muy difícil evaluar, simultáneamente, ítems de diferente tipo. Este formato puede provocar cierto desinterés y desmotivación en los alumnos, llegando incluso a aburrirles. Por esta razón, es deseable que los sistemas de tests computerizados aprovechen sus capacidades multimedia para mejorar las interfaces de los ítems, pero sin que esto suponga ninguna alteración de las propiedades psicométricas de éstos.

Recientemente, algunos investigadores sobre la TRI han utilizado diferentes formatos de ítems en tests adaptativos (Osterlind, 1998). Como se ha mencionado anteriormente, SIETTE incorpora los *siettlets*, que permiten incluir ejercicios con interfaces sofisticadas. El problema de este tipo de ítems, tal y como se ha puesto de manifiesto, es que requieren un esfuerzo notable de desarrollo por parte del profesor. Por este motivo, se ha implementado una biblioteca de plantillas para la construcción automática de tipos de problemas (Guzmán y Conejo, 2004c). La finalidad de esta biblioteca es que los profesores puedan disponer de ejercicios con interfaces visuales más atractivas, desde el punto de vista del alumno, que a su vez se puedan incluir fácilmente en los tests. Los ítems generados utilizando la biblioteca, se evalúan de igual forma que el resto de ítems de SIETTE, por lo que no se pierde el riguroso mecanismo de evaluación de los tests adaptativos.

Las plantillas han sido diseñadas con el objetivo de ser una colección, lo más completa posible, de los ejercicios que un profesor podría incluir en un test. El uso de esta biblioteca permite incluir, en un mismo test, sus ejercicios con los ítems convencionales en la misma sesión de evaluación. La introducción de estos ítems innovadores en los sistemas de tests ha sido ampliamente discutida. Por ejemplo, Huff y Sireci (2001) señalaron que este tipo de ítems novedosos podían mejorar la validez de la evaluación computerizada. También indican que los ítems de opción múltiple, y en general los de papel y lápiz, son inadecuados para evaluaciones de habilidades como la capacidad de razonamiento, síntesis y evaluación. La

introducción de ítems innovadores, por consiguiente, puede ofrecer una mejor evaluación de conocimiento, aptitudes y habilidades.

Boyle et al. (2002) realizaron un estudio en el que se comparaban ítems innovadores con los ítems tradicionales. El test contenía ítems de fuentes públicas e ítems nuevos. Asimismo, incluía tanto ítems interactivos complejos como otros dicotómicos. Los primeros fueron administrados a través de TRIADS, (del inglés, *Tripartite Interactive Assessment Development System*) (Mackenzie, 1999) un sistema que ha sido por el *Centro para el Desarrollo de Sistemas de Evaluación Interactivos* (en inglés, *Centre for Interactive Assessment Development*) de la Universidad de Derby. En este estudio, los ítems fueron analizados utilizando la TCT y la TRI. Los resultados mostraron que los ítems simples, como los de opción múltiple, tenían menor fiabilidad y menor discriminación que los ítems complejos. En conclusión, los ítems innovadores son más útiles para llevar a cabo evaluaciones precisas.

Las plantillas de la biblioteca de ítems representan una colección de los diferentes tipos de ejercicios que normalmente aparecen en los libros de texto. Estas plantillas han sido implementadas mediante *siettlets*, utilizando applets de Java. Los profesores que deseen utilizar una plantilla de la librería no tienen más que instanciarla. Esta tarea se realiza con la herramienta de autor de SIETTE. De esta forma, a pesar de estar desarrollados con *siettlets*, no es necesario que los profesores que vayan a utilizar una plantilla tengan conocimientos de programación.

A continuación se enumeran las plantillas que ofrece la biblioteca:

- a) **Ejercicios de completar:** Son similares a los de respuesta corta. Como se puede apreciar en la figura 6.8, la diferencia radica en que, en este caso, el alumno debe llenar uno o más espacios en blanco que aparecen a lo largo del enunciado. Esta plantilla tiene dos versiones: una versión textual y una numérica. En los ítems textuales, la corrección se realiza comprobando si la respuesta (o respuestas) dadas por el alumno satisfacen una expresión regular suministrada y construida por el profesor. Por otro lado, los numéricos, se corrigen mediante una fórmula suministrada por el profesor que permiten establecer un porcentaje de error para la respuesta es correcta. En el ejemplo de la figura 6.8, el alumno ha de llenar los huecos de un verso, completar el título, y por último, indicar el autor de ese verso. Nótese que en el ejemplo los cuadros aparecen completados de forma correcta.
- b) **Ejercicios de ordenación:** Este tipo de ejercicios son análogos a los ítems de ordenación, presentados en la sección 4.5. La principal diferencia reside en que en la forma de ordenar es visual y/o gráfica (se basa en el uso del ratón). Así, los alumnos tienen que ordenar, un conjunto de n elementos mediante operaciones de arrastrar y soltar. Estos elementos pueden ser texto, imágenes o la combinación de ambos. La figura 6.9 (arriba) muestra un ejemplo de este tipo de ejercicios. El objetivo es ordenar un conjunto de imágenes que representa diferentes estilos arquitectónicos. Tal y como se indica en el enunciado, el criterio que se debe seguir para la ordenación es el orden cronológico.
- c) **Ejercicios de inserción en un conjunto:** En este tipo de ejercicios se le muestran a los alumnos varios elementos, y se pide que inserten en un conjunto aquéllos que cumplen una condición expresada en el enunciado. Para ello, tendrán que realizar

Completa el siguiente verso:

"Con diez cañones por banda,
viento en popa, a toda vela,
no corta el mar, sino vuela
un velero bergantín.
Bajel pirata que llaman,
por su bravura, El Temido,
en todo mar conocido
del uno al otro confín."

Extraído de: *Canción del pirata*.
Autor: José de Espronceda.

Figura 6.8: Ejercicio de completar.

operaciones de arrastrar y soltar, mediante el uso del ratón. Los elementos pueden ser texto, imágenes o la combinación de ambos. Este tipo de ejercicios equivalen a ítems de respuesta múltiple con respuestas independientes, en los que cada elemento es susceptible de pertenecer al conjunto. Si pertenece, equivaldría a que es correcto en el ítem de respuesta múltiple. La figura 6.9 (en medio) muestra un ejemplo de este tipo de ejercicios. En este caso los alumnos tienen que insertar, en el conjunto de la derecha, aquellos títulos (situados en la izquierda del ejercicio) que correspondan a obras de William Shakespeare.

- d) **Ejercicios de emparejamiento:** Este tipo de ejercicios son análogos a los ítems de emparejamiento presentados en la sección 4.5. La principal diferencia radica en que los elementos se muestran en dos columnas y pueden ser textos o imágenes, o la combinación de ambos. Los alumnos tienen que enlazar con líneas trazadas, utilizando el ratón, cada elemento de la columna izquierda con su correspondiente de la columna derecha, en función de la relación que se indica en el enunciado del ejercicio. El objetivo del ejemplo de la figura 6.9 (abajo), en el que se muestra en la columna de la izquierda un conjunto de países, que se tienen que hacer corresponder con su capital, situada en la columna de la derecha.

En la actualidad estos ejercicios se evalúan (en función de cómo los configure el profesor) como ítems verdadero/falso, o como ítems de respuesta múltiple con $n!$ opciones de respuesta, siendo n el número de elementos de una de las columnas.

Como se ha ido mostrado, cada tipo de ejercicio que puede ser construido con una plantilla de la biblioteca tiene su equivalente entre los ítems presentados en el modelo del capítulo 4. Por consiguiente, la inclusión de esta librería no ha supuesto modificaciones sustanciales en la arquitectura de SIETTE, en su mecanismo de evaluación adaptativa, ni en el mecanismo de selección de ítems.

Ejercicios de pasatiempos

Para mejorar la presentación de los ítems, SIETTE ofrece a los profesores plantillas de ejercicios de entretenimiento. Aunque en realidad estos ítems no suponen ninguna aportación desde la perspectiva de una evaluación rigurosa, han sido diseñados para que las

Reordena las siguientes imágenes de manera que aparezcan en orden cronológico:

Inserta dentro del recuadro todos los títulos correspondientes a obras de Shakespeare:

WORKS	WILLIAM SHAKESPEARE
La mujer en la luna	Las alegres comadres de Windsor
Himnos de Astrea	Romeo y Julieta
Othello	Sonnetos de Gulling
Hamlet	

Relaciona cada nombre de un científico con el siglo en el que vivió:

Arquimedes	XX A.C.
Leonardo da Vinci	III B.C.
Isaac Newton	XVIII A.C.
Bernard Bolzano	XV A.C.
Albert Einstein	XVII A.C.

Figura 6.9: Ejercicio de ordenación (arriba). Ejercicio de inserción en un conjunto (en medio). Ejercicio de emparejamiento (abajo).

sesiones de tests sean más amenas y atractivas, en especial para tests de autoevaluación. Esta característica convierte a estos ítems en un mero complemento dentro de un test. En general, estos ítems suelen tener factores de discriminación y dificultad bajos, por lo que no contribuyen a determinar el nivel de conocimiento del alumno con gran fiabilidad. Algunas de estas plantillas son:

- Ejercicios de sopa de letras:** En estos ejercicios el alumno tiene que localizar una o más palabras de entre una matriz de letras. Para ello deberá seleccionar, utilizando el ratón, un conjunto de letras alineadas (bien sea vertical, horizontalmente o en diagonal). En el ejemplo de la figura 6.10 (arriba), el alumno tiene que seleccionar apellidos de personajes que intervinieron de alguna forma en la II Guerra Mundial.
- Ejercicios de puzzles:** Se trata del puzzle clásico ideado por Frank Lloyd. En este tipo de ejercicios, los alumnos deben ordenar las piezas de una imagen para descubrir la forma del objeto representado. En el ítem de la figura 6.10 (abajo) los alumnos tienen que ordenar las piezas del puzzle, tras cuya imagen se esconde el escritor William Shakespeare.

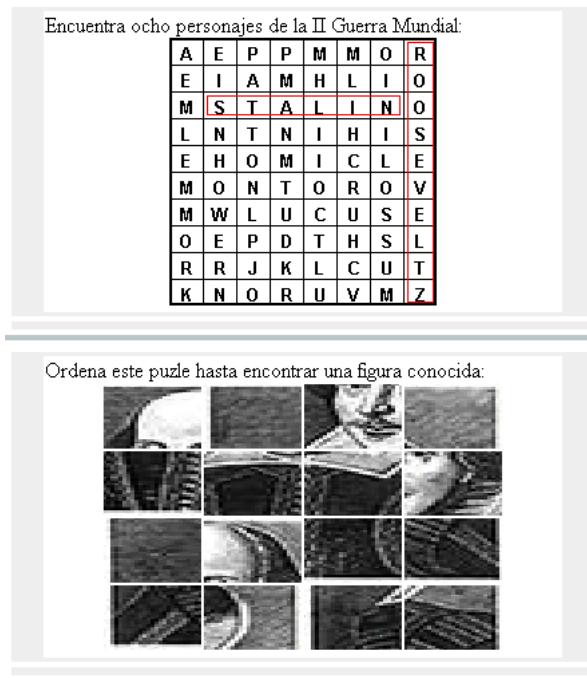


Figura 6.10: Ejercicio de sopa de letras (arriba). Ejercicio de puzzle (abajo).

Tanto los ejercicios de sopa de letras como los de puzzles se evalúan como si se tratara de ítems verdadero/falso. Esto significa que en ambos tipos de ejercicios no cabe la posibilidad de considerar correcciones parciales, es decir, el ejercicio tiene que estar completamente correcto, ya que si no es así, el resultado se considera erróneo.

Para facilitar la generación automática de ejercicios, algunos de las plantillas de ejercicios, definidas en la librería, han sido dotadas de cierto carácter generativo. Los ejercicios de ordenación, los de inclusión en un conjunto y los de correspondencia pueden ser creados (opcionalmente) como ítems con capacidades generativas. En este segundo caso, la idea que subyace es que el profesor incluya en el ejercicio un conjunto con un número considerable de elementos (opciones de respuesta). Asimismo debe indicar cuántos de estos elementos deben mostrarse cada vez que el ítem sea seleccionado. Por lo tanto, en una sesión de test, cuando uno de esos ítems sea seleccionado, el sistema generará un ítem en el que se elegirán, de forma aleatoria, el número predeterminado de elementos de entre el conjunto suministrado inicialmente por el profesor. Por ejemplo, supóngase que un profesor incluye en una de sus asignaturas al ítem representado en la figura 6.9 (abajo). Esto implica que, a través de la herramienta de autor, seleccionará en su momento la plantilla del ejercicio de correspondencia. Además incluirá como combinaciones correctas el siguiente conjunto: *Arquímedes-III d.C., Leonardo da Vinci-XV d.C., Isaac Newton-XVII, Bernard Bolzano-XVIII, Albert Einstein-XX, etc.* Deberá indicar también, que el ítem es generativo, y que

cada vez que se muestre sólo deberían incluirse junto con el enunciado cinco combinaciones. Cuando el motor de inferencia adaptativa seleccione este ítem, se elegirán aleatoriamente cinco pares de elementos, los pares serán separados y ordenados aleatoriamente en dos columnas, y finalmente mostrados al alumno, tal y como se puede apreciar en la figura. El mecanismo de generación para los restantes tipos de ejercicios de la biblioteca es análogo.

6.1.5. Los ítems temporizados

Los modelos psicométricos de respuesta intentan recopilar, a partir de las interacciones con los alumnos, tanta información como sea posible. El objetivo es realizar estimaciones del conocimiento del alumno de la forma más precisa posible. Tal y como se puso de manifiesto en el capítulo 5, desde los comienzos de la psicometría, ha habido diversos intentos de utilizar el tiempo de respuesta, como parte de la información suministrada por el alumno (junto con la propia respuesta al ítem), para la inferencia del valor correspondiente al rasgo latente del alumno.

En SIETTE se ofrece la posibilidad a los profesores de crear y administrar tests o ítems temporizados. Para ello, se almacena internamente el tiempo empleado por el alumno en responder cada ítem. También es posible restringir el tiempo que tienen los alumnos para completar un test o un ítem. A pesar de estas posibilidades que ofrece SIETTE, actualmente, el tiempo de respuesta no se incluye como un parámetro adicional en ninguna de las curvas características de sus ítems, como ya se mencionó en el capítulo anterior, puesto que el modelo de evaluación no utiliza esta información para inferir el nivel de conocimiento del alumno.

Asimismo, aunque, como hacen algunos modelos, podría considerarse que las curvas características difieren dependiendo de si los alumnos han dispuesto de un tiempo límite para realizar el test; en la actualidad no se tiene en cuenta la variabilidad del tiempo de exposición en la calibración de las curvas características.

Las características de temporización de test será estudiadas con más detalle en la sección 6.6.4.

6.1.6. Los ítems externos

SIETTE permite la utilización de ítems externos. Éstos están ubicados fuera de la base de conocimientos, y por tanto, su presentación no está controlada directamente por el sistema, sino que una aplicación externa se encarga de realizarla. Esta aplicación debe tener una interfaz Web, y será normalmente un programa CGI, un servlet, o una página JSP, PHP, etc. Al igual que en el caso de los *siettlets*, estos ítems requieren de uno básico en el que apoyarse. La corrección del ítem se realiza en dos etapas: primero en la aplicación externa y luego en el ítem base.

Estos ítems no están totalmente definidos en la base de conocimientos, como los otros descritos en este capítulo. Así, por cada ítem externo, se almacena la dirección Web donde se encuentra ubicado ese ítem, junto con los parámetros necesarios para invocarlo. Igualmente, se tienen que almacenar el conjunto de posibles respuestas que ese ítem devolverá a SIETTE, así como todas las propiedades psicométricas del mismo; es decir, sus CCO.

Cuando, durante la realización de un test, se seleccione un ítem externo, para presentarlo al alumno se invocará la dirección web de ese ítem con los parámetros necesarios. Como consecuencia, SIETTE pierde, temporalmente, el control de la sesión de evaluación, en favor

del ítem externo (o el sistema en el que éste se encuentre situado). Una vez que el alumno termine de resolver el ítem, deberá ser el propio ítem el encargado de devolver el control a SIETTE a través de una dirección web que SIETTE le envío en la llamada de invocación. Asimismo, a través del parámetro "answer", el ítem externo debe indicar a SIETTE el nombre del parámetro a través del cual se envían las opciones de respuestas que el alumno ha seleccionado. Una vez que SIETTE recupera el control, lee las respuestas, y las evalúa como si se tratara de cualquier otro tipo de ítem. Los ítems externos pueden combinarse en un mismo test con cualquier otro tipo de ítems.

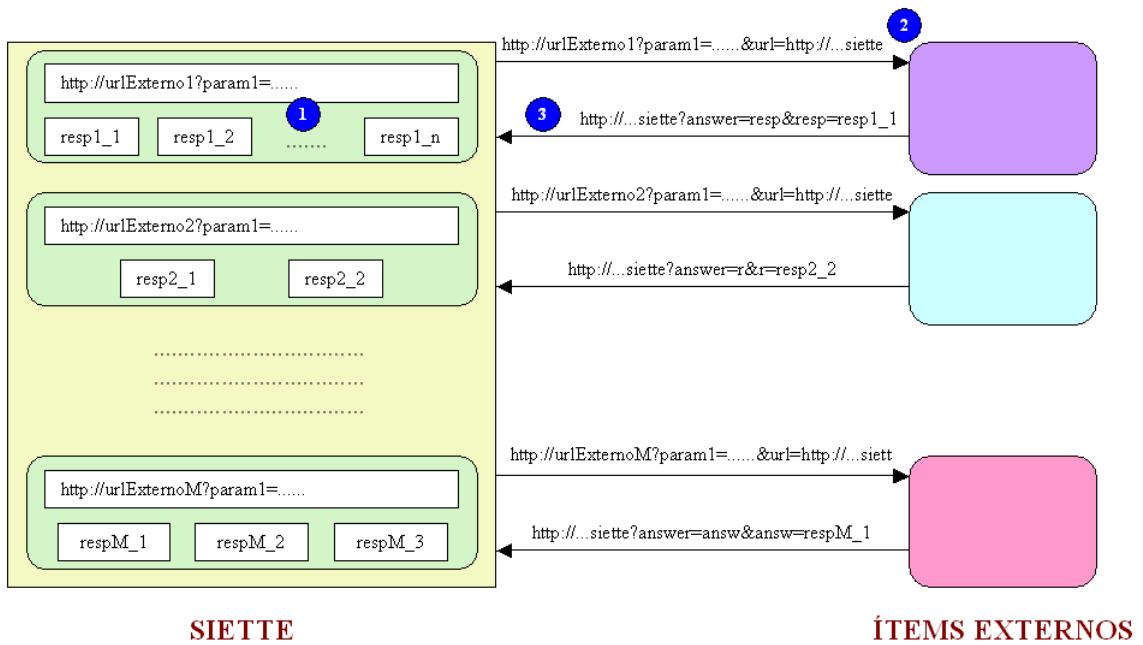


Figura 6.11: Esquema de funcionamiento de un ítem externo.

En la figura 6.11 se ha representado el mecanismo de funcionamiento de un ítem externo. La parte izquierda de la figura representa al sistema SIETTE en el que, como se puede apreciar, se ha incluido M ítems externos. En la parte derecha de la figura se han representado, en diversos colores, los M ítems externos. En la figura se muestra como SIETTE sólo almacena la dirección web del ítem externo, y el conjunto de respuestas (1 en la figura). El conjunto de flechas intermedio, representa el flujo de llamadas que se establecerían entre SIETTE y los ítems externos durante un test en el que se seleccionaran esos ítems. Como se puede apreciar, SIETTE invoca a la dirección web correspondiente, añadiéndole además, a través del parámetro "url" la dirección a la que debe invocar el ítem externo una vez que ha finalizado la evaluación (2 en la figura). En la llamada de vuelta, el ítem externo le indica a SIETTE a través del parámetro "answer", el nombre del parámetro en el que le envía las respuestas que ha seleccionado el alumno (3 en la figura).

Para finalizar esta sección, la figura 6.12 muestra un resumen de los tipos de ítems que ofrece SIETTE. Por cada uno, se ha establecido la correspondencia con el tipo de ítem del modelo de diagnóstico cognitivo al que implementa. Asimismo, se indican las características que cada uno puede poseer, esto es, si puede ser un *siettlet*, generativo, temporizado y/o

ítems de SIETTE	ítem del modelo	sietlet	temporiz.	generativo	externo
Verdadero/falso	Verdadero/falso	*	*	*	*
Opción múltiple	Opción múltiple	*	*	*	*
Respuesta múltiple con opciones independientes	Respuesta múltiple con opciones independientes	*	*	*	*
Respuesta múltiple con opciones dependientes	Respuesta múltiple con opciones dependientes	*	*	*	*
Ordenación	Ordenación	*	*	*	*
Relación	Relación	*	*	*	*
Respuesta corta	Opción múltiple	*	*	*	*
Ejercicio de completar	Opción múltiple	X	*	*	
Ejercicio de ordenación	Opción múltiple	X	*	*	
Ejercicio de inserción	Opción múltiple	X	*	*	
Ejercicio de emparejamiento	Opción múltiple	X	*	*	
Ejercicio de sopa de letras	Verdadero/falso	X	*		
Ejercicio de puzzle	Verdadero/falso	X	*		

Figura 6.12: Tipos de ítems disponibles en SIETTE.

Leyenda: (*) indica que el ítem puede poseer esa propiedad; (X) señala que ese ítem presenta siempre la funcionalidad correspondiente.

externo.

6.2. La base de conocimientos

Esta parte de la arquitectura de SIETTE en la que se almacena físicamente el modulo experto del modelo de evaluación cognitiva presentado en el capítulo anterior. La base de conocimientos está estructurada en asignaturas. Cada asignatura viene representada mediante su árbol curricular de conceptos (su modelo conceptual), tal y como muestra la figura 5.2. Cada concepto tiene asignado un conjunto de ítems que permiten evaluarlo y evaluar también aquellos conceptos relacionados con él, tal y como se estableció en el modelo de evaluación cognitiva. De igual forma, en la base de conocimientos se almacenan las especificaciones de los tests sobre los conceptos del modelo conceptual de la asignatura, así como el conjunto de CCO de cada ítem. Recuérdese que cada opción de respuesta de cada ítem tiene asociada una curva característica por cada concepto que el ítem evalúa. Las asignaturas y sus componentes (conceptos, ítems y tests) se añaden y actualizan a través del editor de tests. Asimismo las CCO se calculan y actualizan mediante los procesos de calibración llevados a cabo en la herramienta de calibración de ítems.

6.3. El repositorio de modelos del alumno

Este repositorio almacena los modelos de los alumnos que han realizado tests a través de SIETTE. El módulo de evaluación actualiza estos modelos durante su realización. Como consecuencia, los profesores pueden visualizarlos a través del analizador de resultados que será descrito en la sección 6.7. Igualmente, la herramienta de calibración hace uso de la información contenida en este repositorio para llevar a cabo la calibración de las CCO de los ítems.

Como se puede apreciar en la figura 6.13, cada modelo del repositorio almacena los vectores con las distribuciones de conocimiento del alumno en todos y cada uno de los conceptos

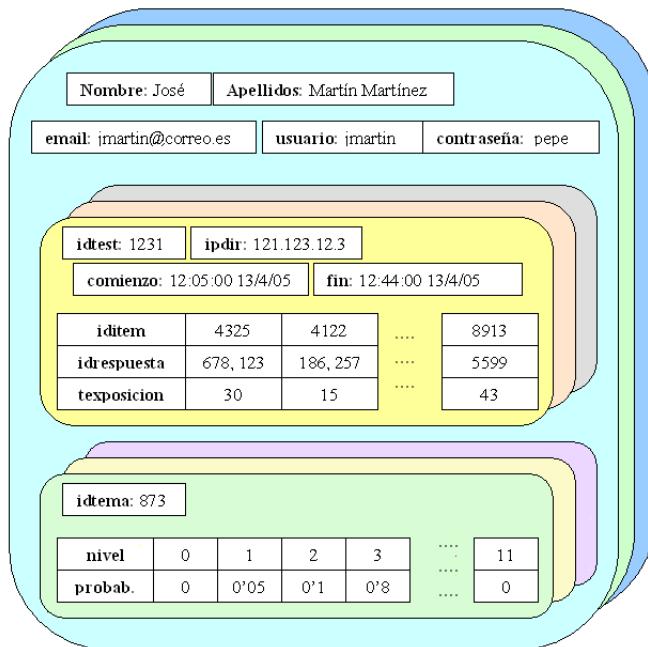


Figura 6.13: Repositorio de modelos del alumno en SIETTE.

evaluados en los tests que ha realizado. Asimismo, por cada test se guarda información para la calibración de las CCO: los ítems que le fueron administrados, en qué orden, qué patrón de respuesta seleccionó, y otros datos menos relevante desde el punto de vista psicométrico (la dirección IP de la máquina desde la que realizó el test, la fecha y hora de comienzo del test y el tiempo empleado en responder a cada ítem). Los profesores pueden eliminar los modelos de aquellos alumnos que hayan realizado alguno de los tests de sus asignaturas, mediante el analizador de resultados.

6.4. El repositorio de profesores

Contiene la información que necesita el sistema sobre los profesores que lo utilizan, tal y como se muestra en la figura 6.14. Almacena el nombre completo; el de usuario y su contraseña, mediante los cuales pueden acceder al editor de tests, al analizador de resultados y al calibrador de ítems; y su correo electrónico.

Un profesor puede tener acceso a más de una asignatura. Cada una tendrá asignada un único profesor que posee el rol de creador; es decir, es aquél que creó inicialmente la asignatura, y que tiene asignado, de forma automática, el nivel máximo de privilegios. Además, puede conceder permisos a otros usuario a través del editor de tests. Las funcionalidades que se ofrecen a un profesor en SIETTE se estructuran en tres partes, correspondientes a los elementos del modelo experto, esto es, temas, ítems y tests, y se estructuran en niveles. Los privilegios se enumeran a continuación, en orden creciente: a) Para los temas: (i) lectura del currículo, (ii) modificación del currículo; b) para los ítems: (i) lectura de ítems, (ii) modificación de ítems; y c) para los tests: (i) lectura de tests, (ii) modificación de tests, (iii)

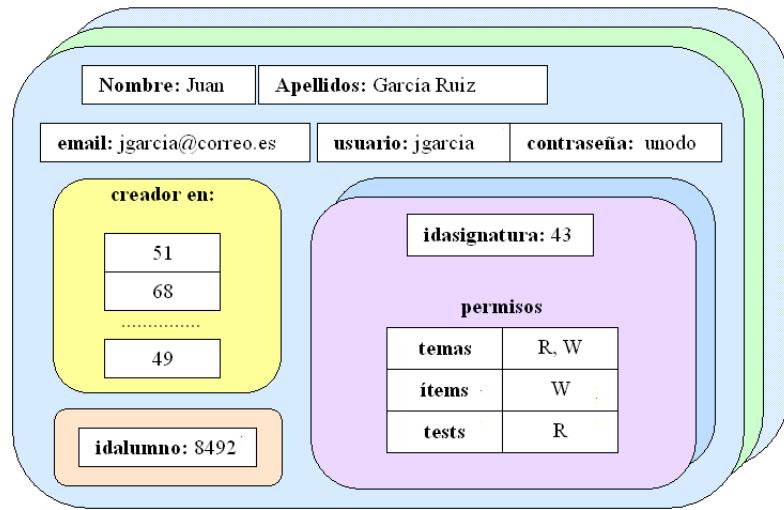


Figura 6.14: Repositorio de profesores en SIETTE.

creación de tests. El usuario *administrador* de SIETTE tiene acceso a toda la información del sistema.

Cada profesor tiene asignada, de forma simultánea, una cuenta de alumno en el aula virtual, con la posibilidad de ver los tests no activos o restringidos a un determinado grupo de alumnos. Ciertamente, esta característica es sólo aplicable a aquellos tests sobre los que el profesor tenga al menos el permiso de lectura.

6.5. El aula virtual

El aula virtual es el lugar donde los alumnos pueden realizar tests. La primera vez que un alumno accede, deberá registrarse con un identificador (o nombre de usuario) y una contraseña. Adicionalmente, puede añadir su nombre completo y su dirección de correo electrónico. Si el alumno así lo solicita, se envía la contraseña a la dirección electrónica almacenada junto al identificador correspondiente.

Una vez registrado, el alumno accede a la lista de asignaturas activas en SIETTE. Tras seleccionar la asignatura, aparece la lista de tests activos en ese momento, para esa asignatura. En la lista de tests, junto con cada test se muestra una descripción sobre su contenido. Una vez seleccionado, se muestra una página con información sobre sus características: número máximo y mínimo de preguntas que pueden aparecer, criterio que se utilizará para evaluar al alumno, número de preguntas que se le mostrarán por página, y finalmente, los temas que comprende el test.

En SIETTE los tests pueden ser de libre acceso o estar restringidos a un determinado grupo de alumnos. La decisión de si un test es de libre acceso la toma el profesor, y la establece durante la fase de elaboración del test. En general, los tests de libre acceso son tests de autoevaluación que sirven al alumno, por tanto, para formarse por sí sólo. Se ha creado una asignatura denominada "Demo" que contiene una colección de tests de libre

acceso. El propósito de estos tests es mostrar las características de SIETTE, en especial, los diferentes tipos de ítems que ofrece, por lo que no tienen ningún valor pedagógico.

La figura 6.2 (izquierda) muestra el aspecto de un ítem de respuesta múltiple en SIETTE. En la parte superior, a la derecha del anagrama de SIETTE, se puede ver el título del test, y debajo de él, el nombre del alumno que está realizando el test. En la parte central de la página, aparece el orden del ítem en el test. Inmediatamente debajo, en un recuadro gris, se muestra el enunciado del ítem. Debajo de este recuadro, una tabla (también de color gris) muestra las opciones de respuesta del ítem. Opcionalmente, el profesor podrá incluir una o más *ayudas*. Estas *ayudas* no son más que pistas que el profesor puede añadir con el objetivo de clarificar el enunciado, y orientar al alumno sobre cuál es la respuesta correcta. En SIETTE, por cada ítem pueden añadirse tantas ayudas como el profesor desee. La inclusión de ayudas se realiza a través del editor de test, durante la construcción del ítem. Así, si el ítem dispone de ayudas y además el test se ha configurado para permitir su aparición, el alumno podrá acceder a las pistas del ítem pulsando un botón de ayuda que aparecerá entre el enunciado y las opciones de respuesta. Si el alumno así lo requiere (pulsando el botón), se seleccionará aleatoriamente una de las ayudas asociadas al ítem, la cual se mostrará en una nueva ventana. Si aún así el alumno lo solicita de nuevo, se volverá a seleccionar otra de las ayudas aleatoriamente hasta que, o bien el alumno envíe su respuesta, o bien no haya más ayudas disponibles. Actualmente, el profesor puede penalizar el uso de las ayudas, en los tests (heurísticos) evaluados por puntos con una puntuación negativa. Una de las líneas de investigación abiertas se basa en el desarrollo de modelos psicométricos más complejos, que permitan que la selección de ayudas sea adaptativa y la evaluación del alumno se vea afectada por las ayudas utilizadas durante el test (Conejo et al., 2003, 2005).

Una vez que el alumno seleccione aquellas opciones de respuesta que crea convenientes, deberá pulsar el botón "Enviar Respuesta". Tras esto, y sólo si se trata de un test de autoevaluación, se mostrará la corrección del ítem. Como se puede apreciar en la figura 6.2 (derecha), el formato de la página de corrección es bastante similar. La única diferencia radica en que las opciones de respuesta correctas seleccionadas por el alumno se señalan con un símbolo de corrección en verde. Las opciones correctas no señaladas por el alumno se marcan con el símbolo de corrección, pero esta vez en rojo. Por último, aquellas opciones señaladas por el alumno pero incorrectas se marcan con una cruz en rojo.

Según Mitrovic (2002), existen evidencias psicológicas de que el *refuerzo* inmediato tras un error, es la acción pedagógica más efectiva; ya que para el alumno, es más fácil localizar y analizar el estado mental que le llevó a ese error e identificar carencias en su conocimiento, que esperar a recibir un refuerzo por parte del profesor. El refuerzo inmediato, a su vez, también reduce la frustración que puede sufrir el alumno debida a carencias en su conocimiento. SIETTE permite la inclusión de refuerzos junto con la corrección del ítem. Los refuerzos se construyen dinámicamente en función de las opciones de respuesta seleccionadas por el alumno.

- En ítems verdadero/falso y en ítems de opción múltiple, cada opción tiene asociada un refuerzo que el profesor incluirá (opcionalmente) a través del editor. En función de la respuesta seleccionada por el alumno, se mostrará, junto con la corrección, el refuerzo correspondiente.
- En los ítems de respuesta múltiple con opciones dependientes, el profesor puede añadir junto con cada patrón de respuestas (combinación de opciones) un refuerzo, si así lo desea. Por consiguiente, cuando el alumno seleccione un patrón de respuestas, y este patrón tiene asociado un refuerzo, en la corrección del ítem se incluirá ese refuerzo.

- El tratamiento de los refuerzo en los ítems de ordenación, es idéntico al caso de ítems de respuesta múltiple con opciones dependientes. Junto con cada ordenación (patrón de respuestas), es posible incluir un refuerzo.
- En los ítems de respuesta múltiple con opciones independientes, cada opción puede tener asociado un refuerzo. Cuando se procede a mostrar la corrección del ítem, el refuerzo que se adjunta equivale a la concatenación de todos los refuerzos de las opciones seleccionadas por el alumno.
- En los ítems de relación se pueden almacenar refuerzos por cada par de elementos de los dos conjuntos que pueden relacionarse entre sí. Así, de forma análoga al caso de los ítems de respuesta múltiple con opciones independientes, el refuerzo que se muestra al alumno equivale a la concatenación de los refuerzos de los pares seleccionados.

En la figura 6.2 (derecha), se puede ver (debajo de las respuestas) el texto del refuerzo suministrado para el patrón de respuesta seleccionado por el alumno.

Tanto ayudas como refuerzos son también fragmentos de código HTML. Por este motivo, al igual que sucede con los enunciados y opciones de respuesta, éstos pueden contener todos aquellos elementos que puede ser incluido en una página HTML.

Desde el punto de vista psicométrico, el uso de refuerzos en TAI puede incumplir una de las hipótesis de la TRI, puesto que con ellos el alumno podría aprender, y por lo tanto, su nivel de conocimiento se vería modificado en el transcurso del test. En la actualidad existen modelos psicométricos (Embreton, 1991) que incluyen la posibilidad de que el alumno aprenda. Ésta es otra línea de investigación abierta.

Una vez que se satisface el criterio de finalización del test, se muestra la estimación final que sobre el nivel de conocimiento del alumno ha realizado el sistema. Por cada concepto directamente evaluado en el test, se muestra un histograma y una tabla con la distribución del conocimiento del alumno. Esta distribución se expresa en la escala de niveles de conocimiento que haya predeterminado el profesor para ese test. Junto con cada distribución de conocimiento se indica cuál es el nivel inferido. Asimismo, el alumno puede ver su distribución y su nivel de conocimiento en todos aquellos conceptos evaluados indirectamente hacia abajo en el test, según la jerarquía curricular del dominio. Obviamente, la información anterior sólo se muestra cuando se trata de un test no basado en heurísticos. Para los basados en heurísticos, por cada concepto, se muestra el nivel de conocimiento estimado heurísticaamente, en función del criterio de evaluación utilizado. En la sección 6.6.3, se explicará cómo se calcula esta estimación heurística.

Además de la información anterior, por cada concepto e indistintamente del tipo de test, se muestra el número de ítems que han sido administrados, y cuántos han sido respondidos correctamente. Si el profesor lo desea, puede configurar el test para que al final se incluya la corrección global, es decir, los ítems que han sido administrados y sus correcciones. La figura 6.15 representa la página generada por SIETTE con la calificación obtenida por un alumno en un test. Debajo de la tabla de resultados aparece un gráfico circular en el que se indica la proporción de ítems por concepto.

Los ítems en SIETTE pueden mostrarse uno a uno (como en la figura 6.2), o bien en grupos de *testlets*, en un número configurable por el profesor. En este último caso, a la hora de decidir qué conjunto de ítems va a ser mostrado al alumno, se aplica el criterio de selección de forma sucesiva hasta que se hayan elegido tantos como tamaño tenga el *testlet*.

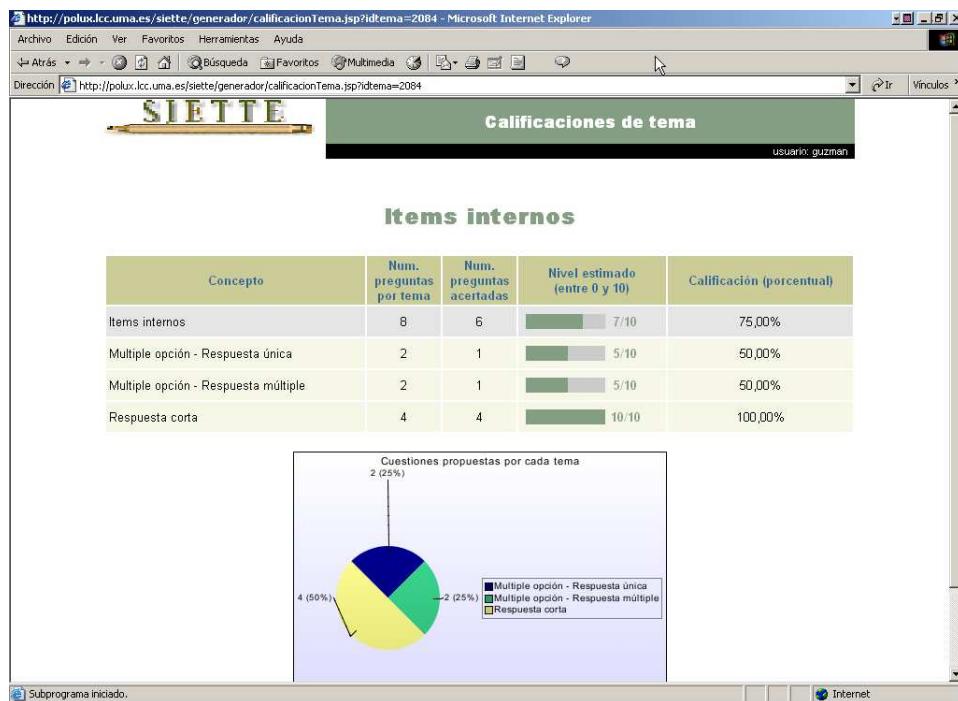


Figura 6.15: Estimaciones del nivel de conocimiento del alumno.

6.6. El editor de tests

Para poder acceder a esta herramienta (Guzmán et al., 2005), los profesores deben introducir su nombre de usuario y contraseña. El administrador del sistema es el responsable de suministrar esta información, y de dar de alta a los profesores en SIETTE. Una vez autenticado en el sistema, al profesor se le muestra la lista de asignaturas que, o bien ha creado, o bien tiene privilegios de acceso. Asimismo, un profesor podrá crear nuevas asignaturas.

6.6.1. Construcción del currículo

La figura 6.16 muestra una imagen del editor tras la selección de una determinada asignatura (en concreto, Java). En la imagen se puede apreciar que la aplicación tiene dos marcos principales. En el marco de la izquierda, se muestra el currículo en forma de árbol. Este árbol puede verse, a petición del profesor, bajo dos modalidades diferentes: *ítems* o *tests*. El cambio de modalidad es tan simple como hacer clic sobre la etiqueta correspondiente, en la parte superior del marco izquierdo. Cuando la opción *ítems* está seleccionada, el árbol muestra la jerarquía curricular de la asignatura estructurada en temas e ítems. Los temas están representados mediante carpetas y los ítems mediante bolas de colores. En función del tipo de ítem, el color de la bola es diferente. Cuando la opción *tests* está seleccionada, el árbol muestra los tests que se han definido para esa asignatura. Bajo cada test, se muestra la estructura curricular de los temas que éste evalúa. Por último, el aspecto del marco de la derecha cambiará en función del tipo de elemento que esté seleccionado en el marco de

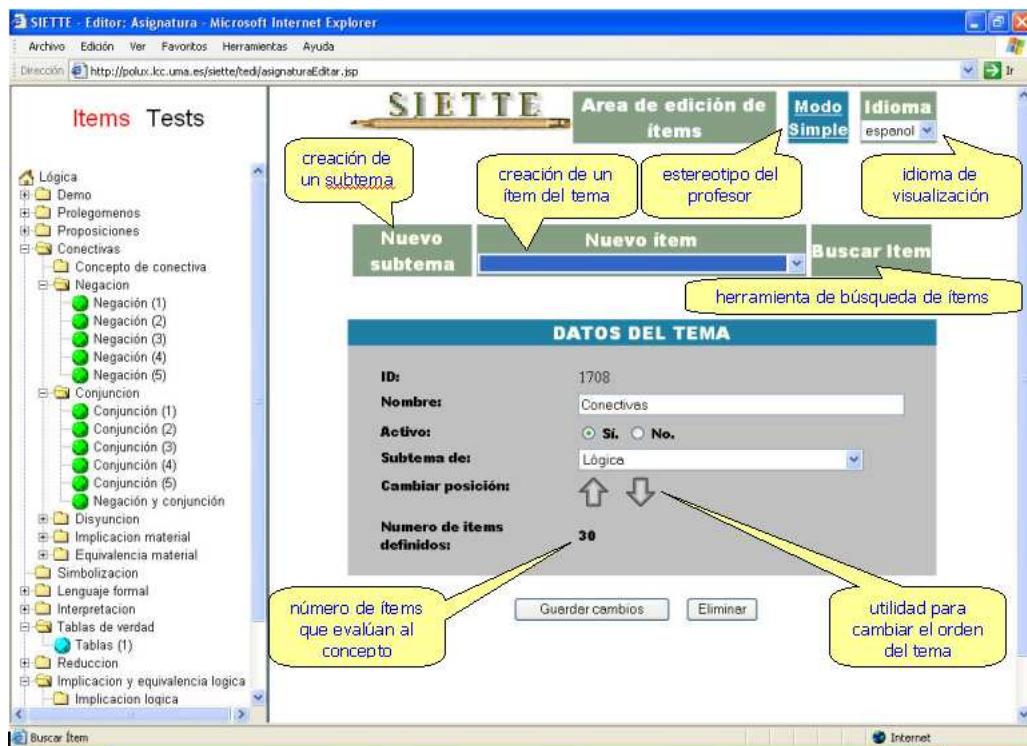


Figura 6.16: El editor de tests: construcción del currículo.

la izquierda. Como resultado, tanto temas como tests e ítems pueden crearse, modificarse o borrarse a través de este marco.

En función del perfil del profesor, el editor puede mostrar los parámetros de los elementos del currículo de dos formas diferentes, y adaptará la presentación de los parámetros en función del perfil actual del profesor. SIETTE tiene dos estereotipos (Kay, 2000) diferentes de profesores: novato y experto. En función de su maestría en su uso, los profesores pueden seleccionar un modo u otro, siendo posible cambiar éste en cualquier momento. La diferencia entre ambos reside en el nivel de detalle de la información que se muestra. En el perfil novato ("modo simple"), parte de la información se oculta. Cuando un profesor con este perfil está editando un elemento, algunos de los parámetros de configuración del elemento asumirán unos valores por defecto. Estos valores por defecto han sido predeterminados por el administrador del sistema. Por otra parte, el perfil experto ha sido concebido para profesores con una cierta experiencia en el uso del sistema, especialmente en lo que a la creación de tests adaptativos se refiere.

El editor es también una herramienta multiusuario. Varios profesores pueden colaborar en la elaboración de los contenidos de una determinada asignatura. En función de sus permisos, el editor adapta la interfaz que se le muestra al profesor. Esto se realiza mediante el mecanismo de ocultación y/o inhabilitación de opciones (Brusilovsky, 2001).

El editor de tests posee además soporte multilingüe. Los temas, ítems y tests puedenadirse en diversos idiomas, de forma que los componentes de una misma asignatura puedan estar traducidos a varias lenguas. Así, a través del editor, el profesor puede incluir las

traducciones de los elementos del modelo experto, para aquellos ítems, temas y tests que considere pertinente. Consecuentemente, en función del idioma que seleccione el alumno cuando acceda a SIETTE, se le mostrarán las asignaturas y tests en esa lengua. El idioma actual en el que está expresado el elemento editado se muestra en la esquina superior derecha de la ventana (6.16). Para editarlo en otro idioma basta con seleccionar la lengua correspondiente en el menú desplegable.

Mediante el editor del currículo es posible construir (de forma visual) el mapa conceptual de una asignatura. Asimismo, a través una utilidad accesible desde esta herramienta se permite localizar uno o más ítems en función de patrones de búsqueda, o incluso crear uno nuevo.

6.6.2. Creación de ítems

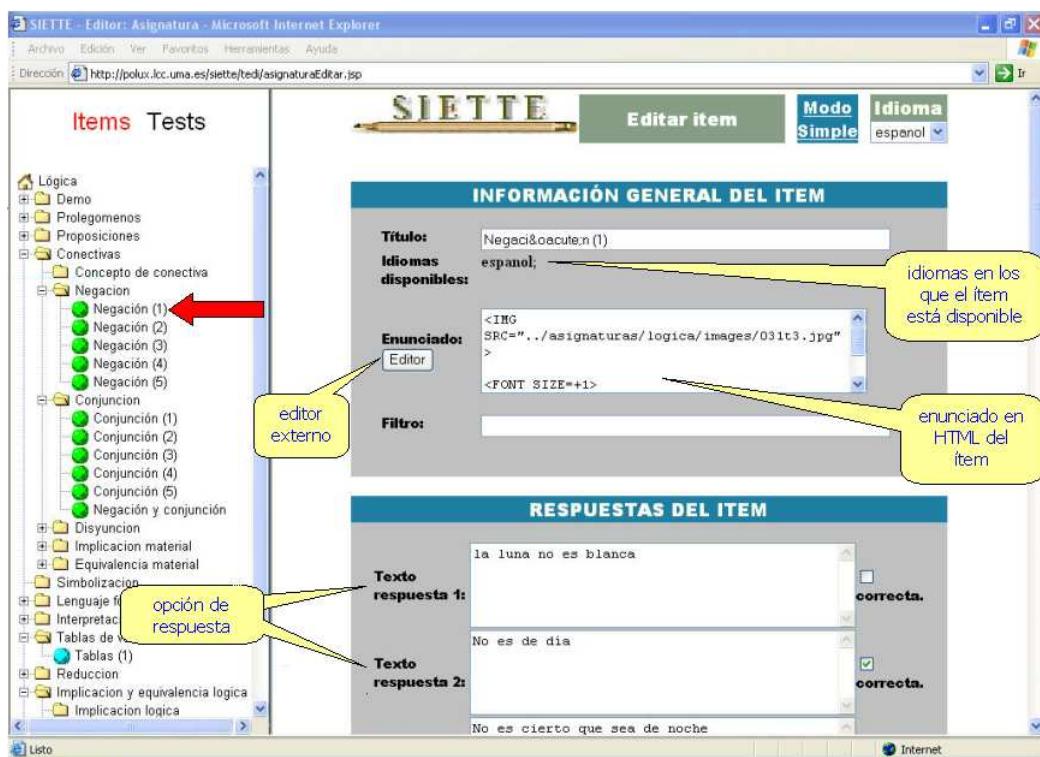


Figura 6.17: El editor de tests: creación de un ítem.

La figura 6.17 muestra la utilidad para la creación y actualización de ítems. Mediante ésta se pueden realizar tareas como añadir nuevos; insertar, borrar o modificar opciones de respuesta; añadir ayudas y refuerzos incorporar el mismo en otro idioma; convertirlos en temporizados; etc.

Esta herramienta permite añadir un mismo ítem en diversos idiomas. Desde el punto de vista práctico, las diferentes traducciones de un mismo ítem se consideran diferentes. Esta diferenciación también se mantiene en los aspectos psicométricos. Las CCO de los

ítems en diferentes idiomas se calibran de forma independiente, lo que permite estudiar el comportamiento diferenciado según el idioma. No obstante, se mantiene una relación entre los ítems en uno y otro idioma, de manera que también es posible estudiar sus propiedades psicométricas de forma conjunta.

Se ha mencionado anteriormente que tanto el texto del enunciado, de las opciones de respuesta, las ayudas y refuerzos se expresan en HTML. Para ayudar a profesores inexpertos en este lenguaje, SIETTE proporciona dos editores externos (figura 6.18):

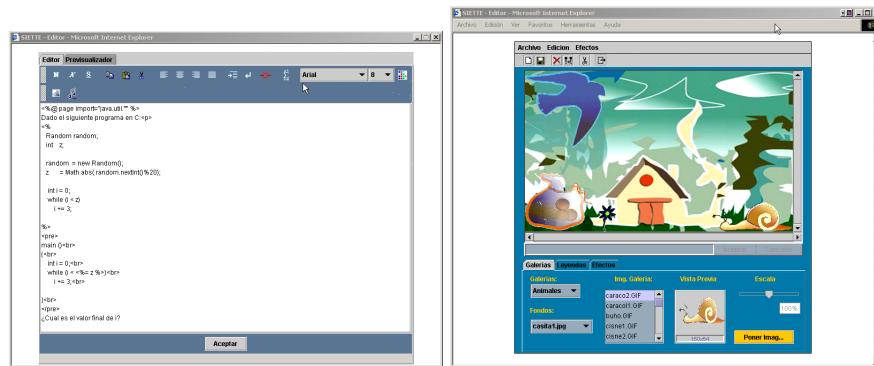


Figura 6.18: Editores externos de SIETTE: Hermes (izquierda) y Edie (derecha).

- Hermes:* Una herramienta visual que permite escribir texto formateado en HTML, e incluso añadir imágenes.
- Edie:* Un editor de escenas que permite construir imágenes mediante la composición de otras y de texto (Román y Conejo, 2003). Esta herramienta ha sido diseñada especialmente para una asignatura de lógica (aunque puede ser utilizada en cualquier otra), que actualmente está en fase de construcción.

Por cada asignatura se configura el editor preferente, de forma que cada vez que se edite una pregunta, aparecerá al lado de los cuadros de texto de todos los elementos editable, un botón que inicia el editor correspondiente. Una vez finalizada la edición en la herramienta externa, el resultado en HTML se incluye automáticamente dentro del cuadro de texto correspondiente en el editor de preguntas de SIETTE.

6.6.3. Definición de tests

A través del editor, los profesores pueden definir tests. En la sección 5.2.1 se explicaron con detalle qué elementos deben configurarse durante la creación de un test adaptativo. Al ser SIETTE un sistema para la generación de tests no sólo adaptativos, el abanico de posibilidades se amplia. La figura 6.19 muestra el aspecto de la utilidad para la edición de tests. Inicialmente, el profesor deberá indicar los temas que van a ser evaluados de forma directa en el test. Otros parámetros de configuración del test se enumeran a continuación:

- *Distribución inicial de conocimiento del alumno:* Antes de haber presentado al alumno pregunta alguna, hay que realizar una estimación a priori de su conocimiento. En

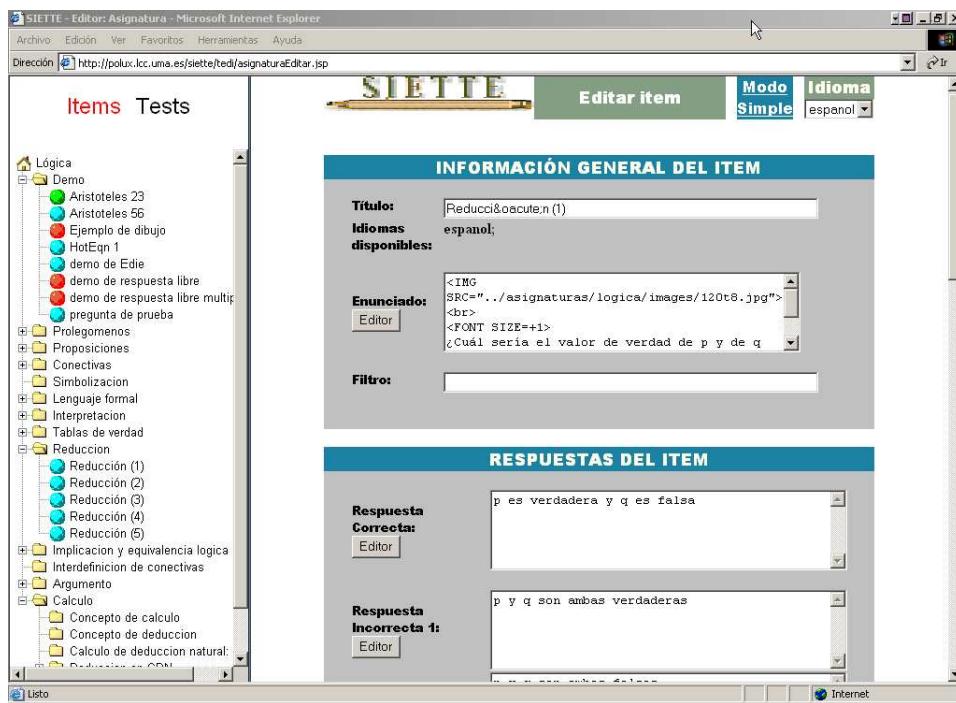


Figura 6.19: El editor de tests: definición de un test.

general, suele asumirse una distribución constante equiprobable, es decir, aquélla en la que todos los niveles de conocimiento tienen asignados la misma probabilidad. Sin embargo, puede darse el caso de que un alumno esté realizando un test de temas sobre los que ya realizó uno o más tests con anterioridad. Como en el repositorio de modelos del alumno se almacena esta información, los tests pueden configurarse para asumir como valores iniciales de conocimiento los resultados obtenidos en tests anteriores. Igualmente, cuando se utiliza SIETTE como módulo de diagnóstico dentro de un sistema instructor, este último puede proporcionarle información sobre el nivel estimado de conocimiento. A partir de esta información, SIETTE construye una distribución normal, centrada en ese nivel estimado.

- *Criterio de selección de ítems:* Además de los criterios adaptativos descritos en la sección 5.3.2 del capítulo anterior, los criterios no adaptativos que pueden utilizarse para elegir el ítem más adecuado, durante la realización del test, son:
 - a) *Aleatorio:* En este caso, los ítems se seleccionan aleatoriamente.
 - b) *En dificultad ascendente:* Antes de comenzar el test, los ítems se ordenan de forma ascendente; así, durante el test, los ítems se elegirán secuencialmente siguiendo el orden anterior.
 - c) *En orden ascendente:* Los ítems se ordenan antes del comienzo del test en orden alfabetico ascendente, según la etiqueta identificativa que acompaña a cada ítem; durante el test, los ítems se eligen secuencialmente.

- *Criterio de evaluación:* Junto con los criterios basados en la TRI, MAP y EAP (explicados en el capítulo anterior), SIETTE permite evaluar a los alumnos siguiendo los dos criterios basados en heurísticos (también explicados en el capítulo anterior): *porcentual* y *por puntos*. Además se incluye el criterio *por puntos con penalización*, que es una modificación del criterio anterior, en el que además de puntuar positivamente los ítems respondidos correctamente, los ítems incorrectamente respondidos puntúan negativamente. Intenta evitar que los alumnos respondan aleatoriamente cuando no conocen la respuesta correcta al ítem. Este criterio también tiene en cuenta la penalización por el uso de ayudas en los ítem. De esta forma, un profesor puede penalizar el uso de ayudas, restándole una cierta puntuación al resultado obtenido en el ítem. A través del editor, se asigna la puntuación que se restará al alumno cada vez que solicite una ayuda. Así, por cada ayuda que solicite en un mismo ítem se le restará esa puntuación.
- *Criterio de finalización:* Los criterios de finalización fueron presentados en el capítulo anterior. Recuérdese que además de los criterios adaptativos, se definieron dos criterios adicionales no adaptativos:
 - a) *Máximo número de ítems:* Se establece un umbral máximo de ítems en el test. En el momento en que el alumno ya ha respondido a ese número de ítems, el test finaliza.
 - b) *Tiempo máximo:* Se establece el tiempo límite que tienen los alumnos para completar el test. Una vez se ha agotado ese tiempo, el test finaliza. La característica de temporización será abordada de forma más detallada en la sección 6.6.4.

Estos dos criterios pueden ser combinados entre sí, de forma que si no se ha llegado al máximo número de ítems, pero ha terminado el tiempo, el test finaliza. Asimismo, estos criterios pueden combinarse con alguno de los criterios adaptativos.

SIETTE permite configurar otras características del test, mediante parámetros:

- *Grupos de alumnos:* SIETTE ofrece la posibilidad de restringir el acceso a ciertos grupos de alumnos predefinidos por algún profesor. Para ello se incluye una utilidad que permite asociar alumnos a un determinado. Un alumno podrá pertenecer de forma simultánea a más de un grupo.
- *Filtros:* Igualmente, es también posible seleccionar, bajo ciertas condiciones, los ítems que pueden formar parte de un test. Por este motivo, se incluye un campo filtro que permite al profesor etiquetar el ítem. Posteriormente, el profesor podrá restringir el banco de ítems de un test a aquellos ítems cuya etiqueta cumpla ciertas condiciones. También es posible filtrar los ítems de un test en función del profesor (o profesores) que haya construido los ítems. Esto permite que múltiples profesores puedan ir añadiendo ítems propios al currículo de una asignatura, y luego cada profesor pueda definir tests con los suyos.
- *Control de acceso:* Al ser SIETTE de un sistema a través de Internet, se le ha dotado de un mecanismo que controla la cadencia de acceso a los tests. Como consecuencia, los profesores pueden configurar sus tests de forma que no se pueda volver a repetir un mismo test, hasta que no haya transcurrido un determinado intervalo de tiempo. A este intervalo de tiempo se le denomina *cadencia*.

El control de la cadencia puede hacerse de dos formas: por alumno y/o por dirección IP. El primero implica que un individuo que acaba de finalizar un test, no podrá volver a repetirlo hasta que transcurra el tiempo indicado en la cadencia. De forma análoga, en la cadencia por IP, el acceso al test se controla por IP. En este caso, cuando se intenta repetir un test desde un PC, independientemente de que el alumno sea distinto, no se permite su acceso hasta que no transcurra el tiempo de cadencia. Esta característica se ha añadido debido a que un individuo podría fácilmente volverse a registrar en SIETTE con un nombre de usuario diferente, y repetir el mismo test.

- *Interrupción de sesiones:* Dados los problemas inherentes de las conexiones a Internet, los tests pueden configurarse para que si un alumno, durante la realización de un test, pierde por algún momento la conexión con SIETTE, pueda volver a reconectarse posteriormente al sistema y continuar el mismo test por donde lo dejó.

6.6.4. Tests temporizados

Como se mencionó en la sección 6.1.5, SIETTE permite temporizar tanto ítems como tests. En cuanto a estos últimos, se pueden definir tres posibilidades:

- *Tests no temporizados:* En este tipo de tests, los alumnos disponen del tiempo que necesiten para completar el test.
- *Tests con ítems temporizados:* Cada ítem tiene asignado un tiempo máximo (indicado durante la fase de construcción). Este tiempo es del que dispone el alumno para responderlo. Si en ese intervalo no se ha enviado la respuesta, automáticamente el generador de tests asume que el alumno deja la respuesta en blanco, y pasa a mostrarle la siguiente cuestión. En un test con ítems temporizados, no todos los ítems tienen que ser temporizados. Sólo aquellos ítems que el profesor indique explícitamente tendrán temporización.
- *Tests temporizados:* En este tipo de tests, los ítems no son temporizados, pero el tiempo requerido para completar el test está restringido. Así, si el tiempo finaliza, se precipita la finalización del test, aún cuando no se satisfaga el criterio de finalización, ni hayan sido mostrados todos los ítems del test.

La característica de temporización de tests e ítems ha sido implementada utilizando un mecanismo similar al empleado en la construcción de los *siettlets*. La figura 6.20 muestra un test temporizado. Como se puede apreciar, el reloj que marca el tiempo del que todavía dispone el alumno para finalizar el test (o el ítem, según sea el caso) se encuentra situado en la esquina superior derecha de la página. Este reloj comienza su cuenta atrás cuando la página con el primer ítem ha sido cargada por el navegador web del alumno. En un test temporizado, cuando el alumno envía la respuesta al ítem (haciendo clic en el botón de enviar), automáticamente el reloj se detiene y el tiempo que marca es almacenado internamente. Una vez que el generador de tests selecciona el siguiente ítem que debe ser administrado, éste se muestra y se reanuda la cuenta atrás en el tiempo correspondiente. Si el reloj finaliza la cuenta atrás y el alumno no ha completado todas los ítems del test, automáticamente se fuerza la finalización, y se muestra la calificación obtenida por el alumno en el test. En cuanto a los ítems temporizados, el mecanismo es análogo. La única diferencia radica en que, una vez que el alumno ha respondido al ítem, el tiempo restante no se almacena, ya que cada ítem tendrá su temporización individual y por tanto un reloj independiente.

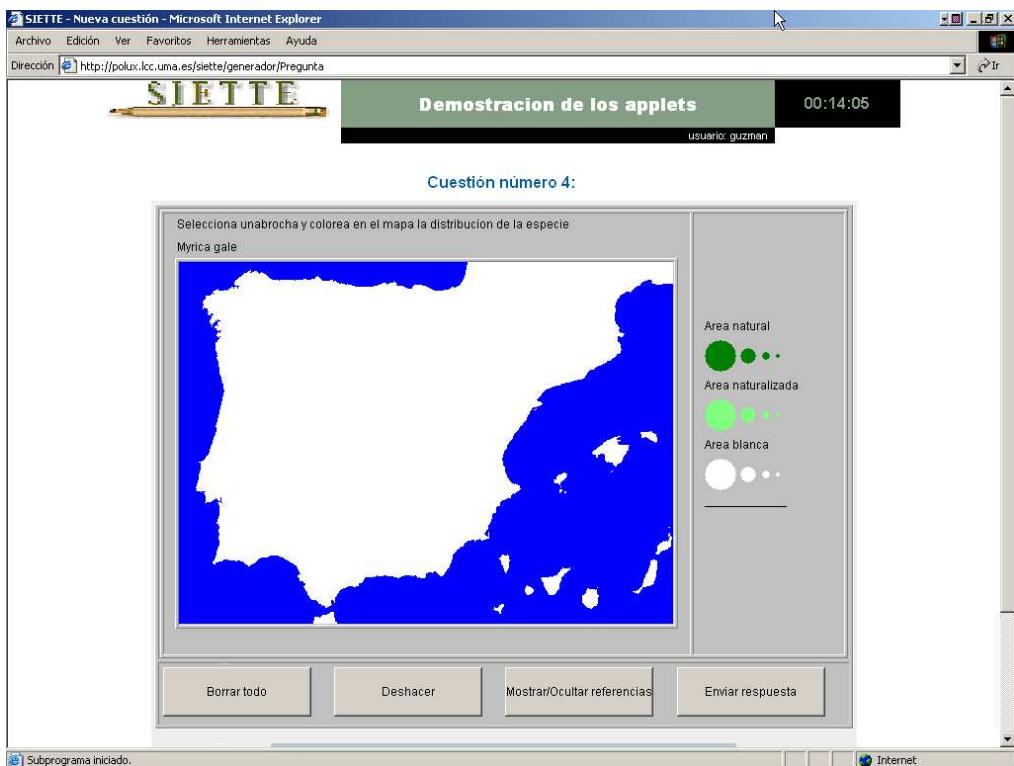


Figura 6.20: Test temporizado.

En general, los profesores podrán decidir qué ítems son temporizados y cuáles no. Para ello, únicamente han de seleccionar, en el editor de tests, la opción de temporización para el ítem (o el test), indicando asimismo el tiempo del que dispone el alumno para responder, expresado en horas, minutos y segundos. Es necesario reseñar que si un test temporizado contiene algún ítem temporizado, la temporización global del test prevalece sobre la del ítem. Consecuentemente, en un test temporizado no se comprueba nunca si los ítems que forman parte de su banco de ítems son temporizado. La temporización de tests o ítems es una característica que se añade al test/ítem dinámicamente. Por este motivo, los profesores pueden, fácilmente y en cualquier momento, configurar los tests/ítems como temporizados o por el contrario, eliminar esta característica, a través del editor.

6.6.5. S-QTI

SIETTE ofrece un mecanismo para la inserción y actualización de contenidos en la base de conocimiento. A través de un fichero en XML con un formato predefinido (descrito con detalle en (Guzmán y Conejo, 2003b)), es posible representar una asignatura completa con su estructuración en temas, ítems y tests. A partir de un fichero de este tipo, pueden insertarse nuevas asignaturas en la base de conocimientos de SIETTE, o bien realizar modificaciones sobre ella, sin necesidad de utilizar el editor de tests. Asimismo, a partir de una asignatura almacenada en SIETTE es posible generar el fichero XML correspondiente en el mismo formato. Además, se ha desarrollado una herramienta (Gálvez et al., 2003) que permite la

conversión de ficheros XML con los contenidos de una asignatura en el formato de SIETTE, denominado S-QTI, a un formato estándar para la creación de tests denominado QTI (del inglés, *Question and Test Interoperability*) (IMS Global Learning Consortium, 2005). Como las características que posee SIETTE son mayores que el conjunto de propiedades que pueden expresarse utilizando el formato que ofrece QTI, las conversiones del formato de SIETTE a QTI están un poco limitadas. Por el contrario, a partir de una especificación mediante QTI se pueden incluir nuevos ítems en SIETTE.

6.7. Analizador de resultados

El repositorio de modelos del alumno almacena información sobre las sesiones de tests realizadas. La herramienta de análisis de resultados de SIETTE permite a los profesores estudiar estos datos. Actualmente, esta herramienta contiene las siguientes funcionalidades:

Ocultar Sesiones	Sesiones de Test	Filtrado Automático							
Realizar Filtrado		Borrar Sesiones Filtradas							
Mostrando 250 sesiones por página.									
Viendo página: 0									
FG	FS	Apellidos, Nombre	Fecha	Duración	Correctas	Hechas	Porcentaje	Puntuación (sobre 100)	Calificación
<input type="checkbox"/>	<input type="checkbox"/>	Guzmán De los Riscos, Eduardo	2005-09-23 12:13:49	0:00	0	0	0	?	0
<input type="checkbox"/>	<input type="checkbox"/>	Demo, Demo	2005-07-22 12:01:39	0:0-2	0	2	0	-8,33	0
<input type="checkbox"/>	<input type="checkbox"/>	Demo, Demo	2005-07-15 13:57:27	0:0-2	0	3	0	-13,89	-1
<input type="checkbox"/>	<input type="checkbox"/>	Demo, Demo	2005-06-30 14:46:58	0:0-2	1	3	33	25,00	2
<input type="checkbox"/>	<input type="checkbox"/>	Guzmán De los Riscos, Eduardo	2005-06-30 11:41:11	0:0-2	1	3	33	36,11	3
<input type="checkbox"/>	<input type="checkbox"/>	Conejo, Ricardo	2005-06-20 03:08:16	0:00	0	0	0	?	0
<input type="checkbox"/>	<input type="checkbox"/>	Conejo, Ricardo	2005-06-09 15:00:56	0:0-2	1	3	33	19,44	2
<input type="checkbox"/>	<input type="checkbox"/>	Conejo, Ricardo	2005-06-09 14:55:54	0:0-2	2	3	66	75,00	8
<input type="checkbox"/>	<input type="checkbox"/>	Guzmán De los Riscos, Eduardo	2005-06-09 14:41:36	0:0-2	0	3	0	8,33	0
<input type="checkbox"/>	<input type="checkbox"/>	Guzmán De los	2005-06-	- - -	-	-	-	- - -	- - -

Figura 6.21: La herramienta de análisis de sesiones.

- *La herramienta de análisis de sesiones:* A través de esta herramienta (figura 6.21) los profesores pueden ver, por cada test, la lista de alumnos que lo han realizado. Para

cada uno, se muestra la siguiente información: el identificador único de la sesión en la base de conocimiento, el nombre completo, la fecha y hora en la que comenzó el test, el número total de ítems que le fueron suministrados, el número de ítems respondidos correctamente y, por último, la calificación que obtuvo en el test. Esta herramienta permite también visualizar cada sesión de forma detallada. Es posible ver todos ítems que le fueron administrados al alumno en el test, el orden en el que fueron mostrados, la respuesta (o respuestas) que seleccionó, y cuál es el patrón de respuestas correcto. Asimismo se dan estadísticas sobre los resultados obtenidos por el alumno. Por cada concepto evaluado se muestra: el nivel de conocimiento estimado del alumno en el concepto, el número de ítems administrados y el número de ítems respondidos correctamente. Finalmente, para cada concepto puede verse una representación gráfica de las distribuciones del conocimiento del alumno. La herramienta permite también eliminar de forma permanente aquellas sesiones que el profesor seleccione.

- *La herramienta de análisis estadístico de ítems:* Suministra información estadística a nivel de ítem, a partir de las sesiones en las que éste ha estado involucrado. Para ello, se selecciona el ítem y el concepto correspondiente. A continuación se muestra una tabla con tantas filas como niveles de conocimiento tenga la asignatura (figura 6.22 (arriba)), donde las columnas representan las posibles opciones de respuestas, y cada fila representa un nivel de conocimiento. Cada celda ij de la tabla (siendo i la fila y j la columna correspondiente) representa el número total de alumnos cuyo nivel de conocimiento, tras haber realizado un test en el que participaba este ítem, sea i , y que a la vez hayan seleccionado como respuesta la opción j .

Frecuencias (Absolutas)		Frecuencias (Porcentajes)		Histograma		Graficas								
Nivel		0	1	2	3	4	5	6	7	8	9	10	11	Suma
<input checked="" type="checkbox"/> r1		1	0	1	4	2	5	3	2	2	0	0	23	43
<input checked="" type="checkbox"/> r2		2	0	0	1	0	0	0	0	0	0	0	0	3
<input checked="" type="checkbox"/> r3		8	0	1	4	1	1	0	2	1	0	0	0	18
<input checked="" type="checkbox"/> r4		9	0	0	2	0	1	2	0	0	0	0	0	14
<input checked="" type="checkbox"/> r5		0	0	0	0	0	0	0	0	0	0	0	0	0
Suma		20	0	2	11	3	7	5	4	3	0	0	23	78

Frecuencias (Absolutas)		Frecuencias (Porcentajes)		Histograma		Graficas								
Nivel		0	1	2	3	4	5	6	7	8	9	10	11	Suma
<input checked="" type="checkbox"/> r1		0,050	0,000	0,500	0,364	0,667	0,714	0,600	0,500	0,667	0,000	0,000	1,000	0,422
<input checked="" type="checkbox"/> r2		0,100	0,000	0,000	0,091	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,016
<input checked="" type="checkbox"/> r3		0,400	0,000	0,500	0,364	0,333	0,143	0,000	0,500	0,333	0,000	0,000	0,000	0,214
<input checked="" type="checkbox"/> r4		0,450	0,000	0,000	0,182	0,000	0,143	0,400	0,000	0,000	0,000	0,000	0,000	0,098
<input checked="" type="checkbox"/> r5		0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000

Figura 6.22: Dos vistas diferentes de la herramienta de análisis estadístico de ítems.

Además, esta herramienta permite mostrar estadísticas acumuladas ((figura 6.22 (abajo))). Es decir, a partir de la información almacenada en el repositorio de modelos del alumno, se muestra una tabla análoga a la anterior, en la que los datos representados son las probabilidades de que un alumno seleccione una determinada opción de respuesta del ítem, dado su nivel de conocimiento estimado por SIETTE. Esta información es relevante para poder llevar a cabo la calibración de las curvas características de los ítems.

6.8. Calibrador de ítems

Gracias a la posibilidad que ofrece SIETTE de realizar tests heurísticos, se puede llevar a cabo la calibración automática de las curvas características de los ítems. Con este fin se ha incluido una herramienta para la calibración de las CCO, que aplica el suavizado núcleo a grupos de ítems del modelo conceptual de una asignatura. Permite la selección del parámetro de suavizado que se desea aplicar al algoritmo de suavizado, así como el conjunto de conceptos del modelo conceptual cuyos ítems serán calibrados. También filtra los resultados de las sesiones que serán utilizados en la calibración de los ítems.

Una vez finalizado el proceso de calibración, esta herramienta visualiza las CCO resultantes. Se puede decidir si se va a volver a aplicar el algoritmo de calibración con otros parámetros, o si por el contrario, las CCO resultantes deben almacenarse en la base de conocimientos de SIETTE.

6.9. La interfaz para las conexiones externas

El objetivo inicial con el que fue ideado SIETTE es servir de módulo de diagnóstico en STI sobre la web. Por este motivo, dentro de su arquitectura (figura 6.26), se ha definido un módulo que permite conectarlo con este tipo de sistemas. Gracias a los mecanismos de evaluación bien fundamentados que proporciona SIETTE, los STI pueden actualizar sus modelos del alumno: la información que poseen sobre el nivel de conocimiento del alumno en los conceptos del modelo del dominio. Como se puso de manifiesto en capítulos anteriores, esta actualización puede llevarse a cabo al principio de la instrucción, cuando el modelo del alumno no contiene ningún valor, lo que correspondería a lo que se definió en el primer capítulo como *evaluación predictiva*. Como SIETTE permite la evaluación simultánea de múltiples conceptos, el STI podría proponerle al alumno la realización de un pretest en SIETTE sobre todos los conceptos del currículo, y utilizar los resultados de ese test para dar valores iniciales al modelo del alumno. Esto permite que la adaptación realizada por el tutor pueda llevarse a cabo desde el comienzo de la instrucción. Igualmente, cada vez que el alumno termine de estudiar un determinado concepto, su nivel de conocimiento podría inferirse realizando un test sobre ese concepto. Consecuentemente, a través de este test se estaría llevando a cabo una *evaluación formativa* del alumno. El planificador de instrucción del STI podría utilizar la información resultante del test para decidir si el estado de conocimiento del alumno en ese concepto es satisfactorio como para proceder a estudiar otro nuevo concepto, o si por el contrario, debe repetir el mismo concepto. Finalmente, al terminar con éste, un post-test realizado en SIETTE permitiría proporcionar al propio alumno una visión global sobre su conocimiento en todos los conceptos estudiados. Esta fase final supone la realización de una *evaluación sumativa* el alumno.

Para permitir la integración en STI, se ha definido un protocolo (Guzmán y Conejo, 2003a) para que ésta sea lo más transparente posible. Es decir, lo deseable es que, el alumno no perciba que el entorno de evaluación es un sistema diferente de aquél donde se está llevando a cabo la instrucción. De esta forma, una vez que el alumno finalice un test, el control debe ser automáticamente devuelto al sistema tutor. Esta integración puede resumirse, a grandes rasgos, en los siguientes pasos (Guzmán y Conejo, 2002b):

1. Inicialmente, el profesor deberá disponer de un nombre de usuario y contraseña que le permitan acceder al entorno de edición de SIETTE. Este par usuario/contraseña deberán ser facilitados por los administradores de SIETTE.
2. El profesor deberá crear una asignatura, y asociar a ésta un currículo compuesto por temas e ítems.
3. Opcionalmente, el profesor podrá construir un conjunto de tests, donde las sesiones de evaluación estén completamente configuradas.
4. En el sistema de instrucción deberán incluirse enlaces a SIETTE, a través de la dirección web preparada para la recepción de peticiones de conexiones externas. Estas llamadas deberán incluir los parámetros adecuados que permitan saber a SIETTE las características del test que debe mostrarse.
5. Una vez que la evaluación haya finalizado en SIETTE, el control será devuelto al sistema de instrucción a través de una o más direcciones web, las cuales habrán sido especificada en los parámetros de la llamada a SIETTE.

En esta sección se hará especial hincapié en los dos últimos pasos, los cuales se centran en la comunicación entre SIETTE y el sistema de instrucción. Este proceso de interacción se rige según el protocolo que se ha definido. Mediante este protocolo se pueden realizar integraciones con diversos grados de acoplamiento entre SIETTE y un STI. El acoplamiento será mayor cuanto más detallada sea la información que necesite recibir el STI sobre el diagnóstico llevado a cabo en SIETTE.

Mediante los parámetros que el STI envía a SIETTE, se podrá configurar dinámicamente una sesión de evaluación, o invocar a un test previamente construido a través de la herramienta de edición. En esta sección no se entrará en detalle a describir los parámetros que puede recibir SIETTE. El objetivo es únicamente dar una visión general de las posibilidades que SIETTE ofrece. Para información técnica más detallada consultese el informe técnico de investigación (Guzmán y Conejo, 2003a).

En función de las características del modelo del alumno que genera SIETTE, las conexiones que un STI puede establecer con SIETTE pueden ser de dos tipos:

- *Sin estado*: En este tipo de conexiones, cada interacción se considera independiente con respecto a cualquier otra. Es decir, SIETTE asumirá que cada interacción corresponde a un nuevo alumno del que no se va a almacenar ninguna información una vez que la interacción finalice. Por este motivo, o bien el STI le suministrará la información de que disponga sobre los valores iniciales del nivel (o niveles) de conocimiento del alumno, o bien SIETTE asumirá unos valores por defecto (según un test predefinido o los parámetros de entrada que reciba).

- *Con estado:* En este caso, el alumno que se conecta a SIETTE debe estar registrado. Su modelo del alumno en SIETTE se almacena entre las distintas interacciones. Este modelo se inicializará con la información almacenada en el repositorio (si es que la hay, y si no asumiendo los valores por defecto), a menos que al comienzo de la conexión se suministren unos valores iniciales.

Para las integraciones con estado, SIETTE permite el registro automático de un alumno a través del propio STI. Mediante un conjunto de llamadas descritas en el protocolo, el sistema de instrucción puede solicitar el registro de un determinado estudiante sin que éste sea consciente de ello.

Asimismo, desde el punto de vista del procesamiento que realiza el STI a posteriori (tras finalizar el test) con los datos suministrados por SIETTE, las conexiones pueden ser de tres tipos: simples, acopladas y mediante servicios Web.

6.9.1. Integraciones simples

SIETTE devuelve el resultado de la evaluación mediante una dirección Web, sin ningún parámetro adicional. El STI deberá, en el momento de comenzar la interacción con SIETTE, enviarle como parámetros una dirección Web diferente por cada uno de los niveles de conocimiento en que se va a evaluar al alumno. De esta forma, una vez que el test haya concluido, SIETTE redirigirá el control a la dirección Web correspondiente al nivel de conocimiento estimado del alumno. Esas direcciones Web permiten al STI recuperar el control de ejecución, a la vez que le indican el resultado de la evaluación. Por ejemplo, si el sistema instructor pasa a SIETTE inicialmente tres direcciones web de salida (parámetros "url"), SIETTE evaluará al alumno en tres niveles de conocimiento (0, 1 y 2), tal y como muestra la figura 6.23. Si al final del test SIETTE estima que el alumno tiene un nivel de conocimiento igual a 1, el control será redirigido a la segunda dirección web que se recibió como parámetro al comienzo de la sesión.

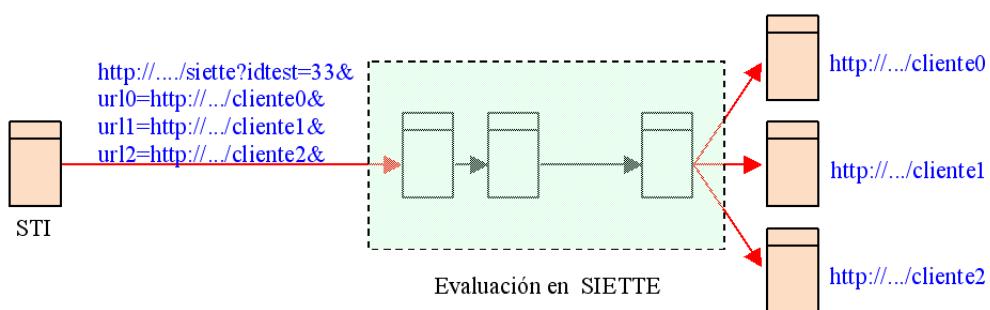


Figura 6.23: Integración simple.

Un ejemplo de integración simple es el test de Piaget³ de habilidades cognitivas mencionado anteriormente en la sección 6.1.2. Ésta fue la primera integración de SIETTE con un sistema externo. En la página inicial de éste, el alumno debe seleccionar el número de niveles de conocimiento en que quiere ser evaluado, entre un máximo de cinco (muy bajo,

³Este test está accesible a través de <http://www.lcc.uma.es/siette/piaget>.

bajo, medio, alto o muy alto). A partir de ahí, y sin necesidad de que el alumno se registre, se procede a la realización del test, que está compuesto por diez *siettlets*. Una vez finalizado el test, el control se redirige a la página correspondiente.

6.9.2. Integraciones acopladas

Este tipo de integraciones son más sofisticadas, ya que requieren un compromiso especial por parte de los STI. Al comienzo de la interacción con SIETTE, el STI le deberá suministrar una dirección web de retorno. Tras finalizar el test, SIETTE enviará, a través de esa dirección, los resultados de la evaluación. Es por tanto necesario que el sistema instructor se involucre de forma más activa en la comunicación con SIETTE, puesto que el STI debe ser capaz de leer y procesar la información que SIETTE le envía. Para el caso de evaluaciones simultáneas de múltiples conceptos ésta es la alternativa más adecuada.

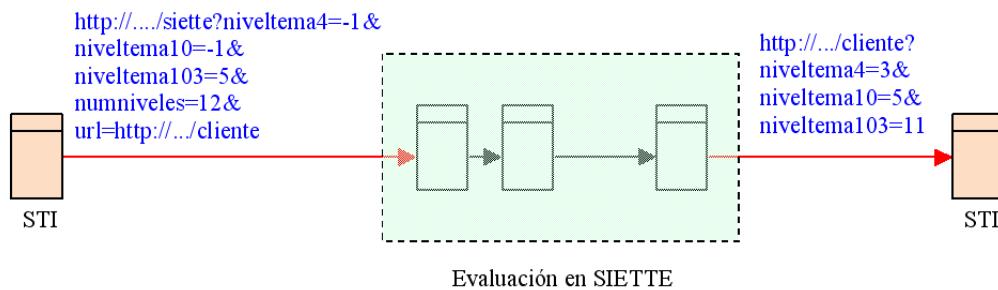


Figura 6.24: Integración acoplada.

La figura 6.24 muestra un esquema de una interacción acoplada entre SIETTE y un STI cliente. Inicialmente, el STI realiza una petición a SIETTE para que inicie un test. Como se puede apreciar, en este test van a evaluarse tres temas (los parámetros etiquetados con *niveltema*). Mediante los parámetros de entrada, se suministran los identificadores de los temas (para el ejemplo: 4, 10 y 103) y el nivel de conocimiento inicial en estos temas (-1, -1 y 5, respectivamente). Nótese que el significado de que un alumno tiene un -1 en un tema es que el STI no tiene ninguna información sobre ese valor, y por lo tanto SIETTE deberá tomar uno por defecto. Igualmente, el parámetro *numniveles* indica el número de niveles de conocimiento en que se evaluará al alumno en el test. Finalmente, mediante del parámetro *url* se informa a SIETTE de la dirección a través de la cual, debe devolver los resultados. Como se puede ver en la figura, el formato en el que se devuelven los niveles de conocimiento estimados, es el mismo en el que se pasan los valores iniciales a SIETTE (para el ejemplo, los valores diagnosticados son 3 para el tema 4, 5 para el tema 10 y 11 para el tema 103).

Otra alternativa a tener que pasar inicialmente los temas que van a evaluarse en el test uno a uno, es la que muestra la figura 6.25. En esta figura se indica directamente el identificador del test (que ha debido crearse con anterioridad a través del editor). Como se puede ver, los resultados se envían de la misma forma que en el caso anterior, con la única diferencia de que, en este caso, SIETTE obtiene la configuración del test a partir su identificador en la base de conocimientos.

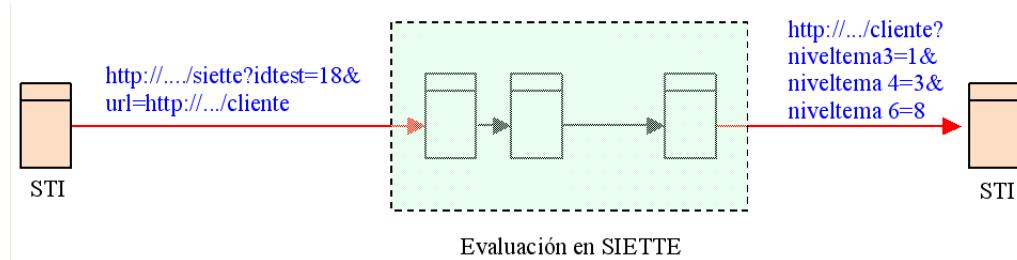


Figura 6.25: Integración acoplada con paso de parámetros simplificado.

Sistemas integrados siguiendo esta aproximación

Utilizando este protocolo de integración acoplada, SIETTE se ha utilizado como módulo de diagnóstico dentro de otros sistemas. Además de en el test de Piaget anteriormente descrito, se ha integrado en sistemas como:

- *TEA (Tutorial de Economía Agraria)*: Es un sistema para el aprendizaje de Economía Agraria (Carmona et al., 2002). Es la versión adaptativa de un sistema estático realizado algunos años atrás. La versión primitiva de TEA (Belmonte et al., 1996a, 1996b), era uno de los primeros sistemas instructores sobre Internet, que además utilizaba tests heurísticos. Para crear la versión adaptativa de TEA⁴ se combinaron dos sistemas: a) SIGUE (Carmona et al., 2002), un sistema de instrucción que permite dotar de características adaptativas a cursos estáticos; y b) SIETTE, a través del cual se proponen tests adaptativos a los alumnos.
- *TAPLI (Tutor Adaptativo de Programación Lineal a través de Internet)*: (Millán et al., 2003, 2004). Se trata de una herramienta para el estudio de Programación Lineal. TAPLI⁵ es el resultado de la integración de un conjunto de componentes en un entorno de aprendizaje. En su arquitectura se pueden distinguir, a grandes rasgos, tres elementos principales: a) Un *componente hipermedia adaptativo*, cuya misión es presentar los conceptos teóricos y los ejemplos; b) un *componente de evaluación*, en el que los alumnos son evaluados sobre los conceptos estudiados en el componente anterior; y c) un *entorno para la realización de ejercicios prácticos*, donde los alumnos pueden hacer ejercicios tutorizados por el propio sistema. Los contenidos del componente hipermedia se estructuran en lecciones (que se han hecho corresponder con conceptos en SIETTE). Asimismo, junto con el material docente de cada lección, se ofrece al alumno la posibilidad de realizar un test para evaluar el grado de asimilación de la misma. El planificador de instrucción de TAPLI recomienda en cada momento qué acción debe realizar el alumno (estudiar una lección, realizar un test, hacer ejercicios, etc.). Cada test contiene un número medio de quince ítems. El alumno es evaluado en los once niveles de conocimiento que se suelen utilizar en España (de 0 a 10). En función del nivel de conocimiento estimado por SIETTE en cada test, TAPLI recomendará la siguiente acción a realizar, aunque en cualquier caso, el alumno será el que decida libremente si quiere continuar con la siguiente lección o repasar la anterior. Es decir, TAPLI es un sistema instructor que aconseja al alumno cuál es el siguiente

⁴El sistema es accesible a través de <http://www.lcc.uma.es/tea>.

⁵Este sistema es accesible a través de <http://www.lcc.uma.es/tapli>.

paso que debe seguir en su proceso de aprendizaje, pero dejándole total libertad. Se elaboraron, utilizando applets de Java, *siettlets* que monitorizan al alumno mientras que realiza ejercicios de aplicación del algoritmo simplex. En función de la destreza del alumno, estos ítems generan ejercicios de mayor o menor complejidad.

- **TRIVIETTE:** El objetivo de este sistema (Machuca y Guzmán, 2004) era crear un entorno que permitiese a varios alumnos jugar de forma colaborativa al trivial a través de Internet. En este sistema, SIETTE desempeña el rol de sistema generador de preguntas. Para diseñar este sistema se extendieron ciertas funcionalidades de SIETTE, permitiendo que diversos alumnos pudieran realizar un test, y que se realizará de forma sincronizada entre ellos. El motor de selección de ítems va eligiendo un ítem que se muestra a todos los alumnos. Sólo aquel jugador al que le corresponda podrá responder. Si falla en su respuesta, únicamente se le mostrará al resto de alumnos la respuesta seleccionada por el alumno anterior. En caso de que ninguno de los alumnos sepa la respuesta correcta, el sistema muestra la corrección. Actualmente, en este sistema, las preguntas se seleccionan siguiendo un criterio aleatorio. El vencedor del juego es aquél que responda más preguntas correctamente. El profesor configura las partidas indicando los temas del currículo de sus asignaturas que formarán parte del test, y el número máximo de preguntas por tema que deben aparecer en el test. La selección del tema se lleva a cabo también aleatoriamente. El sistema también dispone de una utilidad de chat que permite a los alumnos conversar entre sí durante la partida. Actualmente existe un prototipo de este sistema en fase de pruebas.

6.9.3. Integraciones mediante servicios Web

A través de los *servicios Web* (en inglés, *Web services*) los sistemas en Internet pueden comunicarse entre sí de forma completamente transparente al usuario. Esta tecnología se basa en el intercambio de mensajes escritos en lenguaje XML, siguiendo un conjunto de protocolos que determinan, entre otras cosas, el formato de esos mensajes y cómo se pueden comunicar los sistemas entre sí.

SIETTE incluye un servicio Web que le permite comunicarse con STI sobre la Web que implementen esta tecnología. Este tipo de integración representa el mayor grado de acoplamiento entre SIETTE y el STI al que sirve de módulo de diagnóstico.

Mediante servicios Web, SIETTE se integra en los siguientes sistemas:

- **MEDEA:** Se trata de un marco de trabajo para la construcción de STI a través de la Web. El objetivo de MEDEA⁶ (*Metodologías de Enseñanza y Aprendizaje*) (Trella et al., 2002, 2003) es permitir a los profesores reutilizar recursos educativos ubicados en la Web y, a partir de ellos, construir sus propios sistemas de instrucción inteligentes. En MEDEA los componentes se implementan mediante *servicios Web*, y SIETTE es un componente más de evaluación, que los profesores pueden utilizar para incluir tests dentro de sus cursos.
- **LeActiveMath**: Es un sistema tutor inteligente web para enseñar matemáticas⁷⁸. Se construye sobre los cimientos del sistema web ActiveMath (Me-

⁶Existe un prototipo disponible de MEDEA en la siguiente dirección: <http://www.lcc.uma.es/MEDEA>.

⁷El desarrollo de este sistema se enmarca dentro de un proyecto de investigación del VI Programa Marco de la Unión Europea, en el que participan grupos de investigación de diversos países, coordinados por un grupo alemán con sede en el DFKI. Uno de los grupos de investigación que participan en este proyecto es el grupo de *Investigación y Aplicaciones en Inteligencia Artificial (IA²)* de la Universidad de Málaga.

⁸Existe un prototipo de LeActiveMath en la siguiente dirección: <http://www.leactivemath.org>.

lis et al., 2001), descrito en capítulos anteriores. SIETTE está siendo integrado en LeActiveMath, para desempeñar el papel de sistema de evaluación inteligente, facilitando a los alumnos la posibilidad de realizar tests, tanto formativos como sumativos. Igualmente, también será utilizado para la inicialización del modelo del alumno, a través de la administración de pretests. Para este fin, se están extendiendo y aprovechando los mecanismos de integración de SIETTE, que permiten su uso como servicio web. Asimismo, se están utilizando ítems externos, para una mayor flexibilidad en la presentación, y para poder acoplar analizadores algebraicos.

6.10. Arquitectura de SIETTE

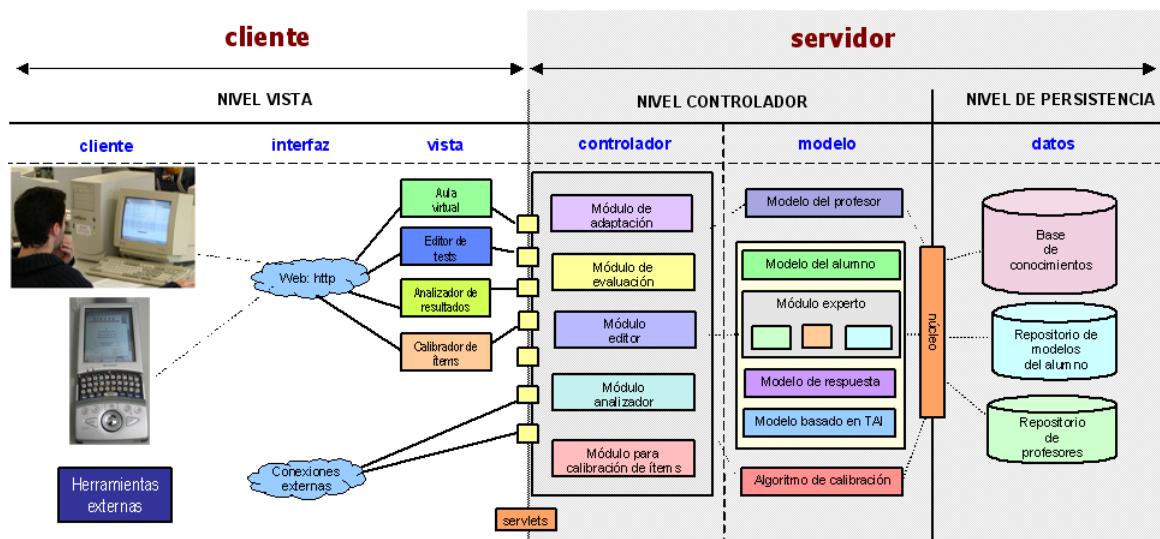


Figura 6.26: Arquitectura del sistema SIETTE.

En la figura 6.26 se ha representado la arquitectura del sistema, que ha sido descrito a un nivel mayor de abstracción en (Guzmán y Conejo, 2004b). Como se puede apreciar, ésta comprende dos secciones principales: la parte del *cliente* y la del *servidor*. A su vez, éstas se dividen en niveles, y éstas últimas se estructuran en capas.

La parte del *cliente* está formada por un único nivel, la *vista*, que a su vez se descompone en tres capas:

- *Cliente*: Es toda aquella herramienta que permite a un usuario, bien sea un profesor o un alumno, conectarse a SIETTE. Pueden ser desde un simple navegador web instalado en un PC o una PDA, a un sistema externo que haga uso de SIETTE como módulo de diagnóstico.
- *Interfaz*: Hay dos diferentes: la de *conexión directa* y la de *conexiones externas*. A través de la primera de ellas los alumnos pueden acceder a SIETTE para hacer tests y los profesores construirlos y analizar los resultados obtenidos. Utilizando la *interfaz*

para las conexiones externas, SIETTE puede funcionar como herramienta de diagnóstico en otros sistemas educativos adaptativos. Como se ha descrito anteriormente, para poder proporcionar esta funcionalidad, se ha definido un protocolo de comunicación con diversos grados de compromiso por parte del sistema externo al que se conecta.

- *Vista:* Está formada por todas aquellas aplicaciones que ofrece el sistema a sus usuario, y que han sido descritas en secciones anteriores. Éstas son:

- El *aula virtual*: Es el sitio web donde los alumnos realizan tests para autoevaluarse, y a través del cual los profesores administran tests a los alumnos para evaluar su nivel de conocimiento con fines académicos. Es la implementación del módulo de presentación del modelo de diagnóstico cognitivo propuesto en el capítulo anterior (figura 5.1).
- El *editor de tests*: Permite la actualización de la información contenida en el módulo experto.
- El *analizador de resultados*: Facilita a los profesores el análisis las sesiones de tests realizadas por los alumnos, y estudiar las propiedades psicométricas de los ítems.
- El *calibrador de ítems*: A través de esta herramienta, se pueden calibrar las CCO de los ítems, tomando como entrada las sesiones de alumnos que han realizado tests.

La parte del *servidor* se estructura en dos niveles: el *controlador* y el de *persistencia*. Este último es el lugar en el que se almacenan físicamente los datos que necesita el sistema. Contiene los tres elementos siguientes:

- La *base de conocimientos*: Lugar en el que se almacenan, de forma permanente, los componentes del módulo experto del modelo de diagnóstico cognitivo. Recuérdese que contiene la estructuración curricular de las asignaturas en conceptos, los ítems asociados a estos conceptos, y las especificaciones de los tests.
- El *repositorio de modelos del alumno*: Donde se almacenan los modelos de aquellos alumnos que han realizado tests a través de SIETTE.
- El *repositorio de profesores*: Contiene toda la información de los profesores que pueden acceder al sistema, esto es, datos personales, grupos creados, permisos de acceso, etc.

El nivel *controlador* es la parte central de la arquitectura. Recibe peticiones desde las interfaces del sistema y lleva a cabo las acciones pertinentes en función de los modelos que contiene, resolviendo, de esta forma, estas solicitudes de servicios y devolviendo la respuesta a través de la vista. Se estructura en dos capas: el *controlador* propiamente dicho, y el *modelo*. La primera contiene la parte del servidor de todas las herramientas disponibles para el cliente a través de la capa *vista*. Está formada por los *módulos de adaptación* y *evaluación*, que se corresponden con los del modelo de diagnóstico del capítulo anterior (figura 5.1). Su misión es construir, administrar y evaluar TAI. Además de los dos anteriores, la capa del *controlador* contiene los *módulos analizador* y *para calibración de ítems*, que implementan la parte del servidor de las correspondientes herramientas ubicadas en la capa de la *vista*.

La capa del *modelo* es el lugar en el que se implementan la mayoría de los elementos de la arquitectura para el diagnóstico propuesta en el capítulo anterior, es decir, el *módulo experto*,

el *modelo del alumno*, el de *respuesta* y el de *diagnóstico cognitivo basado en TAI*. Éstos dos últimos inspiran el funcionamiento de los módulos de adaptación y de evaluación de la capa anterior. Además, se incluyen también el *modelo del profesor*, que es una representación de éste en su utilización del sistema, y el *algoritmo de calibración de ítems*.

6.11. Detalles de implementación

SIETTE es una herramienta web que se ha desarrollado en Java siguiendo la especificación Java EE (del inglés, *Java Enterprise Edition*, y anteriormente conocida como J2EE) desarrollada por la empresa SUN Microsystems, para la construcción de aplicaciones distribuidas sobre Internet. Está implementado mediante una combinación de servlets y páginas JSP, con un núcleo de clases que conforman la lógica de negocios (*middleware*) de la aplicación. Como modelo de programación se ha utilizado el patrón de diseño *Modelo-Vista-Controlador* (en inglés, *Model-View-Controller*), usualmente aplicado en el desarrollo de software siguiendo la especificación Java EE. Actualmente, SIETTE reside en un servidor web para aplicaciones Java EE, de libre distribución, *Apache Tomcat*. Tanto la base de conocimientos como el repositorio de modelos del alumno se han implementado sobre una base de datos relacional, utilizando el sistema gestor desarrollado por Oracle.

En la figura 6.27 se han incluido los diagramas en UML de alto nivel de las clases de SIETTE. Para mayor claridad, éstos han sido organizados en tres partes. La primera de ellas contiene las clases que implementan el módulo experto. En la segunda, aquéllas que intervienen en las herramientas empleadas por los alumnos. Por último, la tercera parte muestra las clases que forman parte de las aplicaciones desarrolladas que utilizan los profesores.

En la sección 6.1.1, se han descrito los tres tipos de ítems de respuesta corta actualmente disponibles. Además de éstos, SIETTE permite la inclusión de nuevos tipos. Para ello basta con implementar una clase que lleve a cabo la corrección en el modo deseado, y hacer que ésta herede de una superclase construida con este fin, que lo único que obliga es a implementar un método que determina si una cadena de caracteres pertenece a un cierto patrón.

El principal problema de los *siettlets*, descritos en la sección 6.1.2, es que su desarrollo está restringido a usuarios con ciertos conocimientos de programación. Esto es debido a que será el profesor (o alguien en quien delegue) el que deberá implementar el programa del applet. Para facilitar esta tarea, en el caso de applets de Java, se ha desarrollado una superclase abstracta que implementa la comunicación con Javascript. De esta forma, el applet del programa deberá heredar de la superclase que ha sido definida. Asimismo, dentro de este applet, deberá implementar forzosamente dos métodos (definidos como abstractos en la superclase). El primero de los métodos es el encargado de, tras evaluar la actuación del alumno frente al programa, devolver la opción (u opciones) de respuesta adecuada a SIETTE. El otro método, cuando sea invocado, deberá mostrar automáticamente la resolución del ítem.

En general, el código HTML de una página web es fácilmente accesible desde cualquier navegador. Esto podría suponer inicialmente un problema en los ejercicios construidos utilizando la biblioteca de plantillas (sección 6.1.4), ya que el código de inclusión del applet en la página web contiene los parámetros de entrada que éste recibe. Esto implica que el alumno podría adivinar fácilmente cuál es la respuesta correcta inspeccionando el código HTML de la página en la que se le muestra el ítem. Para evitar esto, una posible solución es

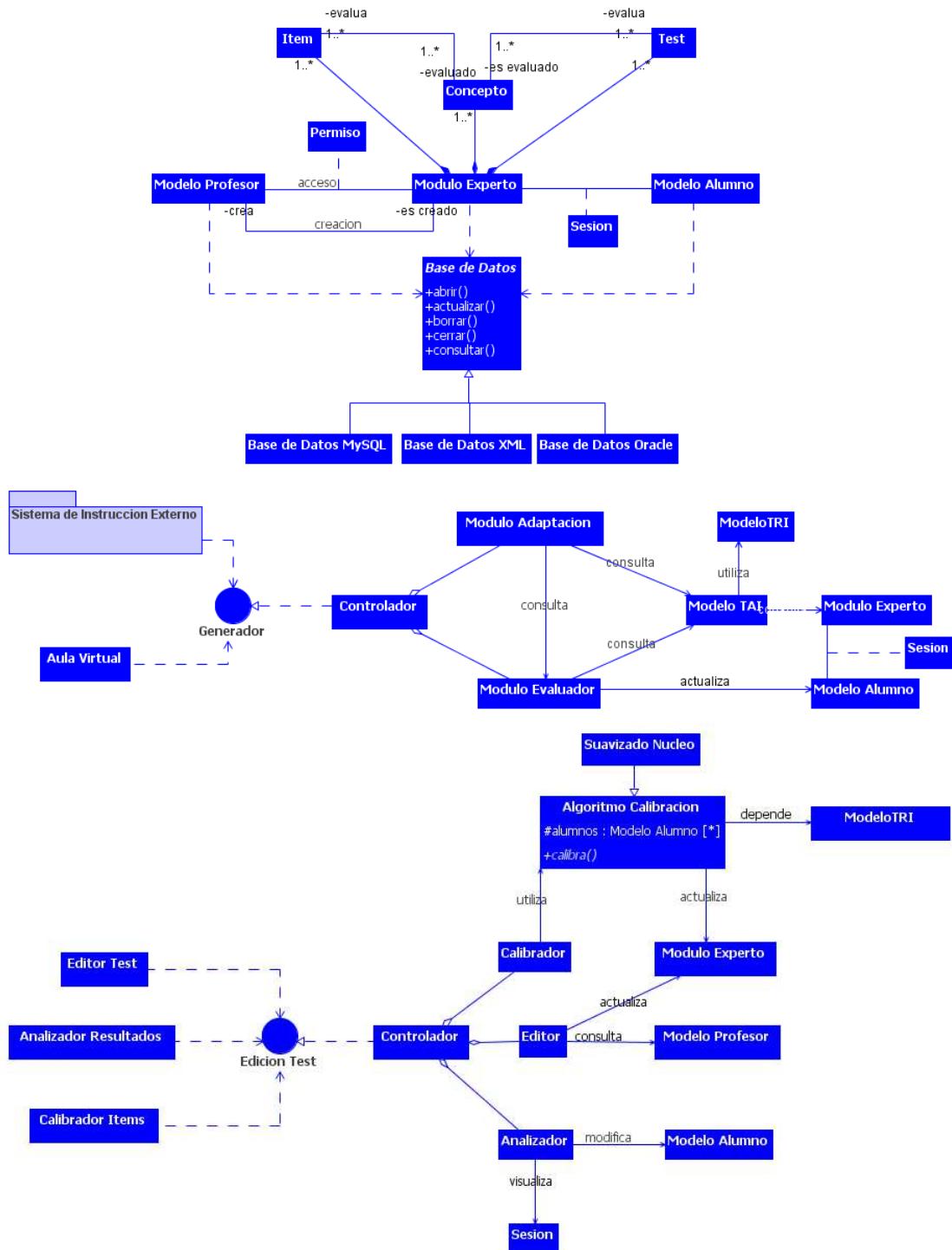


Figura 6.27: Diagramas de clases en UML de SIETTE.

hacer que, antes de que el ítem sea mostrado al alumno, las respuestas sufran un proceso de codificación. Esta codificación sería incluida en la página web como parámetros del applet. De esta forma, dentro del applet, durante el proceso de inicialización, las respuestas codificadas deberían ser decodificadas. Este proceso de codificación/decodificación se podría automatizar y, por consiguiente, la superclase de la que heredan los *siettlets* incluiría todo el soporte necesario para este procesamiento adicional.

En la sección 6.6.2 se ha descrito que SIETTE facilita herramientas externas para editar texto HTML mediante herramientas visuales. Éstos han sido construidos con un mecanismo similar a los *siettlets*. Una clase abstracta (un applet) especifica los métodos necesarios para intercambiar el texto HTML entre SIETTE y la herramienta externa. Para desarrollar una nueva aplicación de edición externa integrable en SIETTE sólo hay que definir una clase principal que herede de la anterior e implemente los métodos de intercambio correspondientes.

6.12. Evolución del sistema

La primera versión del sistema SIETTE fue construida en torno a 1998 como resultado del proyecto de fin de carrera realizado por Antonia Ríos y dirigido de forma conjunta por los doctores José Luis Pérez de la Cruz y Ricardo Conejo (Ríos et al., 1998, 1999). Se trataba de un sistema de generación de TAI en el que ya se utilizaba una discretización de la TRI en la que el número de niveles de conocimiento era fijo e igual a 11. Cada test permitía únicamente evaluar un solo concepto. El modelo TRI utilizado era un modelo dicotómico basado en la función logística 3PL en la que los parámetros de los ítems eran directamente estimados por el profesor. La interfaz del aula virtual se ha mantenido prácticamente igual hasta nuestros días. Este sistema fue implementado en lenguaje C, utilizando la tecnología de CGI combinada con PHP, un lenguaje de programación embebido en HTML para la generación dinámica de páginas web. La base de conocimiento de aquel sistema fue implementada sobre un gestor de base de datos relacionales PostgreSQL.

Posteriormente, se añadieron nuevas funcionalidades y, tomando como referencia la implementación de los TAI que hace SIETTE, se construyó un simulador que permitía la administración de TAI a alumnos simulados en entornos de evaluación controlados. A partir de este simulador, se realizaron diversos estudios empíricos cuyos resultados fueron puestos de manifiesto en (Ríos et al., 1999; Conejo et al., 2000).

En el artículo sobre SIETTE publicado en la revista *International Journal of Artificial Intelligence in Education (IJAIED)* (Conejo et al., 2004) (escrito en torno al año 2000 aunque publicado con posterioridad) se describe en detalle la arquitectura primitiva de SIETTE. Asimismo, en este artículo se vislumbran algunas de las características fundamentales de las que carecía esta versión, y que a su vez ya posee el sistema actual. Como consecuencia, este artículo podría considerarse como un compendio del conjunto de directrices que se establecieron en su momento como requisito indispensable para la nueva versión del sistema SIETTE.

El núcleo actual del sistema SIETTE fue implementado partiendo desde cero, siguiendo un nuevo diseño, una nueva metodología de programación, sobre un lenguaje de programación y un sistema gestor de base de datos completamente diferentes. Este nuevo núcleo comenzó a desarrollarse en la segunda mitad del año 2000. A partir de ese momento se han ido introduciendo mejoras y corrigiendo errores, generándose en consecuencia subsecuentes

versiones del sistema, como consecuencia de la evaluación formativa que ha sufrido el sistema, y que se describirá en el capítulo siguiente. En la actualidad, aunque se dispone de una versión estable del sistema, se siguen realizando nuevas actualizaciones y mejoras. SIETTE se utiliza con cierta asiduidad en la administración de tests de evaluación en asignaturas de diversas titulaciones de carácter universitario, tal y como se pondrá de manifiesto en el siguiente capítulo.

6.13. Conclusiones

En este capítulo se ha presentado el sistema SIETTE, como implementación del modelo de diagnóstico cognitivo propuesto en esta tesis. Este entorno de aplicaciones proporciona a los profesores el soporte necesario para la construcción y administración de tests adaptativos y tests convencionales. En este sentido, incluye un amplio espectro de tipos de ítems que los profesores pueden incluir durante el proceso de elaboración de los currículos de sus asignaturas. Para facilitar esta tarea, se ha dotado al sistema de un mecanismo para la generación automática de ítems isomorfos. Por otro lado, los *siettlets* permiten incluir prácticamente cualquier tipo de ejercicio que pueda ser corregido de forma automática. En esta línea, se ha construido una biblioteca de plantillas de ejercicios. Mediante los ítems externos se han ampliado las posibilidades de evaluación mediante SIETTE, puesto que no es necesario que éstos residan en su base de conocimientos; tan sólo es necesario almacenar cierta información que permita localizarlo a través de la web, e interpretar sus respuestas. Para la construcción y actualización de la base de conocimientos, en la arquitectura de SIETTE, se ha incluido el editor de tests. Esta herramienta posee un conjunto de características que hacen que sus interfaces sean adaptables (Oppermann et al., 1997) al estereotipo de usuario al que pertenezca el profesor y de los permisos que éste tenga.

Los sistemas de generación de tests descritos en el capítulo 2, son herramientas versátiles que, en general, poseen atractivas interfaces. Algunos de estos sistemas incluyen tipos de ítems actualmente no implementados en SIETTE, como los por partes o los de redacción. Por el contrario, SIETTE ofrece otros que no están disponibles en estos sistemas, tales como los ítems de respuesta corta, que son capaces de reconocer un amplio espectro de respuestas, o los *siettlets*. Asimismo, la posibilidad de que los ítems se muestren en conjuntos de testlets es una característica que, en general, no es posible en otros sistemas, y aquéllos que lo incluyen, lo hacen de forma muy limitada. Por último, dentro de esta comparación de SIETTE con otros sistemas, hay que reseñar que SIETTE ofrece mecanismos de evaluación bien fundamentados basados en una teoría consolidada, la TRI, mientras que la gran mayoría de sistemas de generación de tests utilizan heurísticos.

En la actualidad, la base de conocimientos de SIETTE está formada por un total de 77 tablas. Contiene, aproximadamente, 92 asignaturas, 1886 temas, 4146 ítems y 238 tests. Se ha utilizado y/o se utiliza como complemento docente en asignaturas como Procesadores del Lenguaje, Ingeniería del Conocimiento, Ingeniería del Software, Botánica, etc... de titulaciones como Ingeniero en Informática, Ingeniero de Telecomunicaciones, Licenciado en Filosofía de la Universidad de Málaga; y en la E.U. de Ingenieros Forestales de la Universidad Politécnica de Madrid. Asimismo se utiliza también en gran parte de las asignaturas de una titulación de postgrado: el Máster Universitario de Informática Aplicada a las Comunicaciones Móviles, impartido por el departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, y financiado por Fundación Vodafone. Todos estos tests, por motivos de seguridad, están restringidos a usuarios (profesores y alumnos) con unos privilegios determinados.

Además de su funcionamiento como herramienta de tests autónoma, puede incluirse como módulo de diagnóstico en sistemas de enseñanza adaptativos. SIETTE se ha integrado satisfactoriamente en sistemas como TAPLI, y en la actualidad está en fase de integración en otros sistemas como MEDEA y LeActiveMath. Con este fin, se ha desarrollado un protocolo de interacción con otros sistemas que se adapta fácilmente al grado de compromiso requerido por el sistema tutor. Como consecuencia, SIETTE puede utilizarse dentro de otras arquitecturas como herramienta para la inicialización y actualización de los modelos del alumno.

Para poder apreciar algunas de las características de SIETTE, se ha creado una asignatura, "*Demo*", la cual posee una colección de tests en los que se incluyen algunos de los ítems descritos a lo largo de este capítulo.

Parte V

EVALUACIÓN

Capítulo 7

Pruebas y evaluación de la propuesta

Son vanas y están plagadas de errores las ciencias que no han nacido del experimento, madre de toda certidumbre.
Leonardo Da Vinci

7.1. Introducción

En este capítulo se llevará a cabo la evaluación de las aportaciones propuestas en esta tesis. De esta forma, se estudiará la validez, idoneidad y limitaciones del propio proceso de diagnóstico del conocimiento del alumno, según el modelo cognitivo propuesto. Igualmente, se evaluará la herramienta utilizada para acometer esta tarea, esto es el sistema SIETTE. Esto se llevará a cabo bajo dos perspectivas diferentes: una evaluación sumativa de los modelos de respuesta y de diagnóstico cognitivo, y una evaluación formativa de la herramienta SIETTE.

Aunque los conceptos de evaluación formativa y sumativa fueron explicados en el capítulo de introducción, en la siguiente sección, éstos se volverán a abordar desde el punto de vista de la evaluación formal de sistemas y algoritmos.

7.1.1. Evaluación formativa y sumativa

Al igual que sucede en la enseñanza regular, la evaluación de los sistemas educativas puede acometerse desde dos perspectivas: formativa y sumativa. En la *evaluación formativa* (en inglés, *formative evaluation*) se examinan los sistemas que todavía están en fase de desarrollo, con el objetivo de identificar problemas y orientar las posibles modificaciones. Por lo tanto, tiene lugar durante el diseño y las primeras fases de desarrollo de un proyecto, y está orientado a las necesidades de los desarrolladores. Éstos buscan como objetivo mejorar el diseño y comportamiento de sus sistemas. De esta forma, la evaluación formativa puede verse como parte de la propia metodología de construcción de un programa software, y se caracteriza porque su desarrollo se lleva a cabo en ciclos de diseño, implementación y

evaluación formativa. Según Deepwell (2002), la evaluación formativa puede percibirse como un ciclo de *acción-reflexión-acción-reflexión*, y se utiliza para obtener información detallada, que puede emplearse para modificar y mejorar el funcionamiento de un sistema.

Por el contrario, la *evaluación sumativa* (en inglés, *summative evaluation*) se lleva a cabo para cuestionar la construcción, comportamiento o salidas de un determinado sistema. Su objetivo es, por tanto, probar la idoneidad de los formalismos o técnicas utilizadas.

Aunque existen un número reducido de autores, como por ejemplo Winne (1993), que critican la separación entre evaluación sumativa y formativa propuesta por Scriven (1967), y que consideran que ambas son lo mismo; la gran mayoría de investigadores (Mark y Greer, 1993; Shute y Regian, 1993; Barros, 1999) aceptan esta separación en dos tipos de evaluación. Según Littman y Soloway (1988), la evaluación formativa responde a la pregunta: ¿Qué relación existe entre la arquitectura del sistema y su comportamiento? Por el contrario, la sumativa tiende a responder a otra cuestión importante: ¿Cuál es el impacto del sistema en el alumno? Por otra parte, según Shute y Regian (1993), la distinción entre evaluación sumativa y formativa reside en que la segunda se rige según una condición interna de control, y responden a la siguiente cuestión: ¿Cómo se puede mejorar el sistema?; mientras que la primera se rige por condiciones externas de control, y responde a la siguiente pregunta: ¿Cómo es este sistema en comparación con otros sistemas o aproximaciones?

Aunque la evaluación sumativa y la formativa son diferentes, son también complementarias. Gran parte de la información recopilada durante las actividades llevadas a cabo para la evaluación formativa, deben tenerse en cuenta desde el punto de vista de la sumativa, puesto que permiten detectar imprecisiones en los formalismos utilizados.

Los experimentos formales se utilizan principalmente en evaluaciones sumativas, donde el objetivo es evaluar la efectividad de un sistema completo (Twidale, 1993). Las ventajas de este tipo de experimentos son evidentes, y su principal objetivo es medir la efectividad global del sistema cuando éste es utilizado por un gran número de alumnos. La evaluación controlada es la técnica que suministra más información útil sobre la efectividad de un sistema. A su vez, también es la medida más certera, puesto que se ajusta adecuadamente al paradigma científico de objetividad y capacidad de reproducción.

Para la evaluación sumativa del modelo de diagnóstico cognitivo propuesto en esta tesis, durante las primeras secciones de este capítulo, se va realizar un estudio empírico controlado del modelo de diagnóstico que ha sido presentado en el capítulo 5. Este estudio se ha llevado a cabo utilizando una de las estrategias expuestas por Murray (1993) para la evaluación de sistemas inteligentes: simular el comportamiento humano. Por este motivo, se han utilizado los denominados *alumnos simulados*. Éstos son artificios software que intentan emular el comportamiento de examinandos reales cuando se les administra un test. En (Millán, 2000) se justifica el porqué resulta adecuado el uso de alumnos simulados en vez de estudiantes reales, para evaluar el comportamiento de un modelo de diagnóstico. Estas razones se enumeran a continuación:

- No parece adecuado utilizar un modelo de diagnóstico con estudiantes reales sin haber comprobado previamente su funcionamiento correcto con alumnos simulados.
- En caso de utilizar examinandos reales para validar el modelo sería necesario disponer de otro mecanismo de diagnóstico que permitiera llevar a cabo una correlación entre ambos resultados. En general, la forma de realizar la validación es comparar los resultados del modelo con las estimaciones que el profesor realiza tras los exámenes. El inconveniente de esta solución es que las estimaciones de los profesores son heurísticas, con lo que los resultados del estudio de la validez del modelo quedarían en entredicho.

Los resultados obtenidos en cada uno de los experimentos, que se describirán en este capítulo, no son consecuencia de simulaciones aisladas. Por cada uno de los casos, se han llevado a cabo un gran número de simulaciones, y los resultados expresados en las diferentes tablas, corresponden a los valores medios.

Este capítulo comienza con la descripción del simulador que se ha utilizado para efectuar la evaluación sumativa del modelo de diagnóstico. El primer experimento realizado se basa en el estudio de cómo afecta la inclusión de diferentes tipos de ítems al rendimiento y precisión de los TAI. A continuación se lleva a cabo un estudio comparativo de los diversos mecanismos de selección adaptativos incluidos en el modelo. Estos criterios serán comparados entre sí, para determinar cuál de ellos es el más apropiado en cada caso. Posteriormente, se efectuará un análisis comparativo del modelo de respuesta presentado en esta tesis, frente al paradigma más popular, esto es, el modelo de respuesta continuo y dicotómico basado en la función 3PL. En este estudio se realizarán diferentes experimentos, que permitan analizar las diferencias en cuanto a número de ítems, precisión y tiempo de cómputo requerido por los TAI. Otro estudio que se llevará a cabo se basa en probar empíricamente que los criterios de selección de ítems adaptativos, en tests sobre múltiples conceptos, son capaces por sí solos de realizar una elección de ítems balanceada en contenido. Se analizará también el algoritmo de calibración propuesto para el modelo de respuesta. Este estudio estará compuesto por varios experimentos, en los que se evaluará el algoritmo en función de los resultados de muestras de alumnos de diferente tamaño utilizadas para la calibración; se analizará también el valor del parámetro de suavizado más adecuado en función del tamaño muestral; y se pondrá de manifiesto cómo influyen tanto la función núcleo seleccionada, como el criterio de evaluación heurístico utilizado en la primera fase del algoritmo, en la bondad de los resultados de la calibración. Asimismo se comparará la propuesta original del algoritmo de calibración realizada por Ramsay (1991) con las modificaciones que se sugieren en esta nueva versión del algoritmo.

La evaluación de la herramienta de diagnóstico, el sistema SIETTE, se realizará en dos experimentos. En el primero de ellos se describirá un estudio realizado en 2002, en el que se compararon los resultados de calibraciones del mismo banco de ítems, a partir de evidencias recogidas mediante tests de papel y lápiz, y mediante SIETTE. Finalmente, se realizará la evaluación formativa del sistema, mediante la descripción de sus diferentes prototipos, y de los experimentos realizados con cada uno de ellos.

7.2. El simulador

Para llevar a cabo la evaluación sumativa de las aportaciones realizadas en esta tesis, se ha desarrollado un simulador. Éste ha sido implementado a partir del núcleo de clases del sistema SIETTE, y a su vez aprovechando muchas de las funcionalidades implementadas otro simulador realizado anteriormente (Ríos et al., 1999; Conejo et al., 2000). Como resultado, esta herramienta emula el comportamiento del sistema SIETTE a la hora de evaluar estudiantes reales, puesto que intenta recrear sesiones de diagnóstico en un entorno controlado. A grandes rasgos, el simulador, en función de los parámetros de entrada que recibe, genera un determinado número de alumnos simulados, un modelo conceptual con estructura jerárquica, un banco de ítems y un test. Posteriormente, se lleva a cabo la simulación del comportamiento de los alumnos frente a un test, y se muestran los resultados.

7.2.1. Generación del currículo

El simulador inicialmente construye el currículo de la asignatura. Para ello, recibe como entrada una secuencia de números naturales que determinan su estructura jerárquica. De esta forma, si recibiera como parámetro la secuencia: a, b, c , indicaría que el concepto que representa a la asignatura (concepto raíz) tiene a hijos. Cada uno de éstos tiene a su vez b descendientes, y por último, éstos tienen c . Como consecuencia, se construye un árbol completo, cuya profundidad viene determinada por la longitud de la secuencia de entrada. Así, para la secuencia de ejemplo a, b, c , el número total de conceptos del currículo sería igual a $a + a * b + a * b * c + 1$. Conforme se van creando, los conceptos se van etiquetando con número naturales de forma consecutiva a partir de 0, valor que se asigna al raíz.

7.2.2. Generación del banco de ítems

Una vez construido el currículo, se procede a generar el banco de ítems. Se crearán tantos ítems como se indique a través de un parámetro de entrada. El número de preguntas de cada tipo es también configurable y se determina a priori. Cada ítem generado será asignado consecutivamente a uno de los conceptos del currículo. La finalidad es que haya una distribución balanceada de ellos por concepto. Igualmente, cada ítem va a tener asociado un número de opciones de respuesta determinado, según un parámetro de entrada, que diferirá dependiendo del tipo de ítem. Este último se determina mediante una función, que va estableciendo, de forma circular, el tipo del siguiente ítem.

Por cada ítem se generan sus CCO, en función de su tipo, tal y como se describe a continuación:

- Ítems verdadero/falso: Se generan los tres parámetros correspondientes a una función 3PL discreta. La generación de la discriminación, dificultad y adivinanza se lleva a cabo según los parámetros de entrada del simulador. De esta forma, la curva característica discreta se construye a partir de una discretización de la función logística 3PL, para valores naturales de θ entre 0 y el número de niveles de conocimiento menos uno; donde este último valor es también un parámetro de entrada del simulador. Los tres parámetros de la función 3PL se generan de forma automática y análoga para todos los ítems, en función de unos valores de entrada que recibe el simulador. Tanto el factor de discriminación como el de adivinanza pueden fijarse para todas las curvas características de los ítems del banco. Además, el simulador permite que los valores de ambos parámetros sean diferentes por cada CCI. En este caso, se construye una distribución normal centrada en el valor del parámetro de entrada, y según ella, se irán generando aleatoriamente los valores de los parámetros de la CCI de cada ítem. Por último, la dificultad de los ítems se suele generar siguiendo una distribución normal cuyos valores posibles son los mismos que puede tomar el nivel de conocimiento del alumno. Una vez generada la CCI, ésta se convierte en la CCO de la opción de respuesta correcta, y su opuesta en la de la incorrecta.
- Ítems de opción múltiple: En este caso también se genera la CCI global del ítem, siguiendo un procedimiento análogo al que se sigue para las cuestiones verdadero/falso. Posteriormente, la CCO de la respuesta correcta se hace corresponder con la CCI. Las curvas características del resto de opciones (todas ellas incorrectas), se generan de diferente forma, en función de un parámetro de entrada, que determinará si los ítems van a ser dicotómicos o politómicos. Si van a ser dicotómicos, cada CCO se calcula

aplicando la siguiente fórmula: $\frac{1-CCI}{n}$, donde CCI es la curva característica del ítem y n es el número de opciones de respuesta erróneas. Si van a ser politómicos, las curvas de las $n - 1$ primeras respuestas se generan siguiendo la siguiente fórmula, para cada nivel de conocimiento θ :

$$CCO_j(\theta) = \frac{c_j}{1 + e^{-1,7a_j(\theta-b_j)}} \quad (7.1)$$

Los valores de los parámetros a_j , b_j y c_j , para la respuesta j -ésima, se generan siguiendo distribuciones normales similares a las que siguen los parámetros de la CCI. Nótese también que a_j toma siempre valores negativos en la fórmula 7.1, para asegurar que la curva sea monótona decreciente.

Por último, la curva que falta se calcula a partir del resto de CCO. Cada ítem es en sí mismo un espacio probabilístico; por tanto, la suma de las CCO de todas las respuestas debe ser igual a uno. De esta forma, la curva de la respuesta que queda se calcula restándole a una distribución constante igual a uno, la suma de todas las CCO. Ciertamente, en algunos casos, es necesario ajustar los valores de alguna de las CCO generadas según la fórmula 7.1, para evitar que la última de ellas contenga valores negativos.

- Ítems de respuesta múltiple con opciones independientes: Para construir este tipo de ítems, se genera una función 3PL discreta por cada una de las opciones de respuesta, siguiendo el procedimiento aplicado en los dos tipos de ítems anteriores. Si la opción es correcta, a su CCO se le asigna directamente la función 3PL correspondiente. En caso de que sea incorrecta, se le asigna la opuesta de la función 3PL correspondiente. Por este motivo, la CCI global del ítem se calcula aplicando la fórmula 4.10, y su dificultad (necesaria cuando se aplican el método de selección basado en la dificultad) se infiere a partir de la ecuación 5.33.
- Ítems de respuesta múltiple con opciones dependientes: El cálculo de sus CCO es análogo al caso de ítems de opción múltiple. A partir de las opciones de respuesta se determina el número total de CCO que son necesarias. Recuérdese que al tratarse de ítems con opciones dependientes, cada CCO está asociada a un patrón de respuestas. Así, sea m el número de opciones, el número de CCO de este ítem será igual a 2^m . Por esta razón, se asume que este ítem es de opción múltiple con $2^m - 1$ opciones de respuesta, y según esto, sus CCO se calculan siguiendo el procedimiento correspondiente.
- Ítems de ordenación: Éstos son análogos a los de respuesta múltiple con opciones dependientes. Por este motivo, el cálculo de sus CCO es similar. Cada opción de respuesta, en este caso, corresponde a una posible ordenación de los elementos del ítem, habiendo por lo tanto $m!$ opciones posibles.
- Ítems de emparejamiento: Son análogos a los anteriores, por lo que su tratamiento también es equivalente.
- Ítems de asociación: Este tipo de ítems equivalen a los de respuesta múltiple con opciones independientes. Cada opción equivale, en este caso, a determinar si un determinado emparejamiento es correcto. Habrá por lo tanto, tantas CCO como posibles emparejamientos.

Por cada uno de los conceptos evaluados por un ítem (bien sea directa o indirectamente) se repite este proceso hasta obtener todas sus CCO. Asimismo, los ítems del banco podrán estar o no correctamente calibrados, en función de un parámetro de entrada.

A partir del banco de ítems, se construye el test. Éste se configura a partir de un conjunto de parámetros de entrada que, entre otras cosas, indicarán los criterio de selección de ítems, de evaluación y de finalización, así como los umbrales de estos últimos. Además, habrá que indicar qué conceptos del currículo van a ser evaluados directamente.

7.2.3. Generación de los alumnos simulados

Los alumnos simulados se generan en un número igual al parámetro de entrada correspondiente. A cada uno de ellos se le asigna un identificador numérico consecutivo a partir de uno. El comportamiento posterior de cada alumno en el test va a venir determinado por su modelo cognitivo real. Se trata de un modelo de superposición sobre los conceptos evaluados en el test. El nivel asociado a cada concepto lo determina el simulador a priori (antes de comenzar el test), de la siguiente forma: A los conceptos hoja del currículo evaluados en el test, se les asigna un nivel de conocimiento, denominado *nivel de conocimiento real*, generado como un número pseudoaleatorio. Éste sigue una distribución normal, cuyos valores permitidos son aquéllos que pertenecen al rango entre cero y el número de niveles de conocimiento menos uno. Para calcular el nivel de conocimiento real del alumno en los conceptos que preceden a éstos, se le somete a un test en el que se le administran todos los ítems (uno detrás de otro) que evalúan a ese ítem, bien sea directa o indirectamente. En este test se utilizan las CCO reales de los ítems, y el criterio de evaluación aplicado en este caso puede ser el MAP o el EAP. De esta forma, se obtiene como resultado el nivel de conocimiento real en ese concepto. Para los precedentes se aplica el mismo procedimiento de forma análoga.

7.2.4. Administración simulada del test

Tras determinar el modelo cognitivo real de cada alumno simulado, se procede a administrar el test bajo las condiciones indicadas en la simulación. El procedimiento empleado es análogo al que se seguiría para un estudiante real. La única excepción está en el modo en el que el alumno simulado selecciona el patrón de respuesta, una vez que se le ha mostrado un ítem. La respuesta del alumno se simula de la siguiente forma: Sea C_j el concepto que evalúa directamente el ítem mostrado, se genera un número aleatorio r entre 0 y 1. En función del tipo de cuestión, se procede como se indica a continuación:

- Ítems verdadero/falso: Para calcular la opción de respuesta seleccionada, se toma la probabilidad asociada al nivel de conocimiento real del alumno en la CCI del concepto en cuestión. De esta forma, si el nivel de conocimiento real en el concepto es θ^r , si el valor de r es menor o igual que esa probabilidad ($r \leq CCI(\theta^r)$), esto significará que el alumno selecciona la respuesta correcta. En otro caso, habrá elegido la incorrecta.
- Ítems de opción múltiple, de respuesta múltiple con opciones dependientes, de ordenación y de emparejamiento: Se toman las probabilidades de todas las CCR del ítem para el valor correspondiente al nivel de conocimiento real del alumno en el concepto. Estas probabilidades se van acumulando progresivamente en función de una determinada ordenación. Por ejemplo, sea θ_j^r el nivel de conocimiento en el concepto C_j ;

supóngase que el ítem i que se ha mostrado al examinando tiene tres posibles patrones de respuesta (u_{i1}, u_{i2}, u_{i3}), además de la respuesta en blanco (u_{i0}). Considérese además, que las probabilidades asociadas a θ_j^r en cada patrón de respuesta (según la CCR correspondiente) son las siguientes: $P_{\overrightarrow{iu_i1}}(\overrightarrow{u_{i1}}|\theta_j^r) = 0,5$, $P_{\overrightarrow{iu_i2}}(\overrightarrow{u_{i2}}|\theta_j^r) = 0,15$, $P_{\overrightarrow{iu_i3}}(\overrightarrow{u_{i3}}|\theta_j^r) = 0,25$ y $P_{\overrightarrow{iu_i0}}(\overrightarrow{u_{i0}}|\theta_j^r) = 0,1$. Estos valores se acumulan quedando expresados de la siguiente forma: 0,5, 0, 0,65, 0,9 y 1,0. Una vez hecho esto, en función del valor de r , el patrón de respuesta elegido es el siguiente: Si $r \leq 0,5$, entonces es el primero de ellos; si $0,5 < r \leq 0,65$, será el segundo; si $0,65 < r \leq 0,9$, el patrón seleccionado será el tercero; por último, si $r > 0,9$, la elección corresponderá a la respuesta en blanco.

- Ítems de respuesta múltiple con opciones independientes e ítems de asociación: Por cada opción de respuesta, y en función del valor de r , si el valor de la probabilidad asociada al nivel de conocimiento real del alumno en el concepto es menor o igual que esa probabilidad, se asume que el examinando ha elegido esa opción. Si por el contrario es mayor, se asume justo lo contrario. Se procede de esta misma forma, por cada una de las opciones de respuesta, y como resultado se obtiene las opciones seleccionadas (y no seleccionadas), obteniéndose por tanto el patrón de respuesta.

Una vez determinado el patrón de respuesta que selecciona el alumno simulado, se procede a inferir su nivel de conocimiento. Tras esto, se vuelve a elegir el siguiente ítem, según el criterio de selección con el que esté configurado el test. Este proceso se repetirá hasta que se cumpla la condición de terminación del test. Una vez que éste ha finalizado, en función del propósito de la simulación, y por tanto, dependiendo sus los parámetros de entrada, se muestra diversa información que permite estudiar y analizar los resultados obtenidos.

7.2.5. Metodología del análisis

En los estudios que se muestran a continuación, el cómputo estadístico de las pruebas se ha realizado de forma aproximada. En cada experimento, cada una de las pruebas llevadas a cabo se ha repetido un mínimo de diez veces, y los datos expresados en las tablas corresponden a valores medios. En algunos análisis se ha incluido además la desviación típica de los resultados como indicador de su factor de confianza. Así, siendo R el resultado de un determinado experimento y d su desviación típica, el valor verdadero de R estará incluido en el intervalo de confianza $[R - d, R + d]$ con el 95 % de probabilidad. Esta información sólo se detalla en aquellos experimentos en los que este dato es relevante. En el resto de los casos, el valor de la desviación típica es lo suficientemente pequeño para garantizar las hipótesis con el 95 % de certeza.

7.3. Estudio sobre la inclusión de diversos tipos de ítems en el banco de ítems de un test

El objetivo de este estudio es analizar cómo afecta la inclusión de diversos tipos de ítems en un banco utilizado para la administración de TAI. Es decir, se busca determinar si las diferentes clases de preguntas que propone el modelo de respuesta aportan alguna mejora, o por el contrario suponen algún inconveniente en el rendimiento o en los resultados del TAI.

Parám. CCI		100 % OM	75 % OM	50 % OM	25 % OM	0 % OM
Discrim = 1,9	Núm. ítems	6,80	3,63	3,63	3,53	3,32
	Error	0,28	0,20	0,16	0,13	0,18
Discrim = 0,7	Núm. ítems	14,89	13,40	13,34	13,66	12,22
	Error	2,79	0,74	0,86	0,77	0,46
Discrim = 0,5	Núm. ítems	14,94	15,62	15,34	16,23	16,33
	Error	3,85	1,55	1,22	1,15	0,85
Discrim = 1,9 Adivin = 0,25	Núm. ítems	12,08	7,32	7,30	6,78	6,14
	Error	0,48	0,27	0,34	0,34	0,28
Discrim = 0,7 Adivin = 0,25	Núm. ítems	23,29	22,08	21,93	22,95	21,85
	Error	3,26	1,79	1,34	1,31	0,65
Discrim = 0,5 Adivin = 0,25	Núm. ítems	30,18	26,53	27,11	25,68	30,58
	Error	8,07	2,05	2,07	2,39	1,57

Tabla 7.1: Comparación entre el número medio de ítems en una sesión de evaluación, en función del tipo de ítems del banco y de los parámetros con los que fueron generadas sus CCO.

7.3.1. Experimento 1: Comparación según los parámetros de las CCI

En este estudio se sometió a una población de 100 alumnos simulados a un test de un único concepto en el que se disponía de un banco compuesto por 300 ítems. Éste contenía ítems de dos tipos: de opción múltiple (con tres opciones) y de respuesta múltiple con opciones independientes (con cinco opciones), en proporciones diferentes según la simulación. El resto de tipos no se han incluido debido a su equivalencia con uno de estos dos, tal y como se ha puesto de manifiesto en la sección 7.2.2.

El test construido evaluaba a los examinandos en 12 niveles de conocimiento. El nivel de conocimiento real de cada individuo fue generado siguiendo una distribución normal centra en el valor medio de la escala de niveles de conocimiento, es decir, en cinco. El mecanismo de selección utilizado era el método bayesiano de la máxima precisión esperada, y el criterio de finalización el basado en la máxima precisión de las estimaciones, en el que el umbral establecido es igual a 0,001. El nivel de conocimiento se estima utilizando el criterio MAP.

7.3.2. Resultados

La tabla 7.1 muestra los resultados correspondientes a diferentes simulaciones realizadas con bancos, en los que la proporción de ítems de opción múltiple y de respuesta múltiple con opciones independientes era diferente. Asimismo, los parámetros con los que sus CCO fueron generadas, también eran diferentes. Los valores que aparecen en la primera columna fueron utilizados para generar la CCO de la respuesta correcta en los ítems de opción múltiple; y para crear todas las CCO en el caso de ítems de respuesta múltiple. Las curvas se crearon siguiendo el clásico modelo logístico 3PL, donde los tres parámetros se indican precisamente en esa columna. Como se puede apreciar, sólo aparecen los valores correspondientes a la discriminación y la adivinanza. Éstos son realmente los valores centrales de las distribuciones normales que se siguieron para generar las discriminaciones y adivinanzas de las CCO,

respectivamente. La única excepción es la adivinanza de las seis primeras filas de datos, en las que todas las CCO fueron generadas con adivinanza igual a cero. Por ejemplo, para los datos relativos a las dos primeras filas, el valor asociado a la discriminación es igual a 1,9. Esto indica que las discriminaciones utilizadas para generar las CCO fueron a su vez generadas siguiendo una distribución normal centrada en 1,9. Asimismo, el valor asociado a la dificultad se genera siempre siguiendo una distribución normal centrada en el nivel de conocimiento medio. Es decir, si el test evalúa en 12 niveles de conocimiento, el valor medio obtenible por un examinando será 5.

Las columnas de la tercera a la séptima indican la proporción de tipos de ítems presentes en los bancos utilizados para las sesiones de tests. La tercera columna indica que el 100 % de los ítems eran de opción múltiple; la cuarta expresa que tan sólo lo eran el 75 %, y por consiguiente, el 25 % restante de respuesta múltiple con opciones independientes. La quinta columna indica que la mitad de los ítems del banco eran de opción múltiple y la otra mitad de respuesta múltiple con opciones independientes; la sexta señala que el 25 % de ellos eran de opción múltiple. Por último, la séptima indica que todos eran de respuesta múltiple con opciones independientes.

Cada par de filas de datos contiene los resultados correspondientes al número medio de ítems requeridos en las sesiones de evaluación (*Núm. ítems*), y el factor de confianza (*Error*) del correspondiente valor. Esta información fue extraída de las sesiones de evaluación realizadas con bancos de ítems de las características indicadas en cada columna, y cuyas CCO fueron generadas a partir de la función 3PL con los parámetros indicados en la fila correspondiente.

En la figura 7.1 se han representado, de forma comparativa, los valores correspondientes al número medio de ítems requeridos para el diagnóstico del nivel de conocimiento de los alumnos de la muestra. Como se puede apreciar, el eje de abscisas no está escalado, y en él se representan los valores centrales de las distribuciones normales según las cuales se han generado las discriminaciones de las CCO de los ítems. La gráfica superior corresponde al número de ítems requerido para aquellos bancos cuyas CCO se han creado con adivinanza igual a cero, mientras que la inferior representa los resultados con ítems cuyas CCO siguen una distribución normal de adivinanzas centrada en 0,25. Sólo se han plasmado los resultados correspondientes a bancos en los que todos los ítems eran de opción múltiple, frente a aquéllos obtenidos sólo con cuestiones de respuesta múltiple con opciones independientes.

Tal y como se puede apreciar, en la mayoría de los casos, el número de ítems cuando el banco está formado por cuestiones de respuesta múltiple con opciones independientes es menor que si éste está formado exclusivamente por preguntas de opción múltiple. De hecho, cuando las discriminaciones son muy altas, los resultados con ítem de respuesta múltiple implican una reducción en algo menos de la mitad del número de ítems requeridos.

7.3.3. Experimento 2: Comparación según el criterio de selección de ítems

El objetivo de este experimento es estudiar si influye el criterio de selección en el número de ítems requerido para el diagnóstico del conocimiento del alumno, en función además, del tipo de ítems presente en el banco empleado en el test.

Con este fin, se consideraron tres tipos de bancos. El primero de ellos compuesto tan sólo por ítems de opción múltiple. El segundo compuesto, en partes iguales, por ítems de opción múltiple y de respuesta múltiple con opciones independientes. El último de ellos estaba

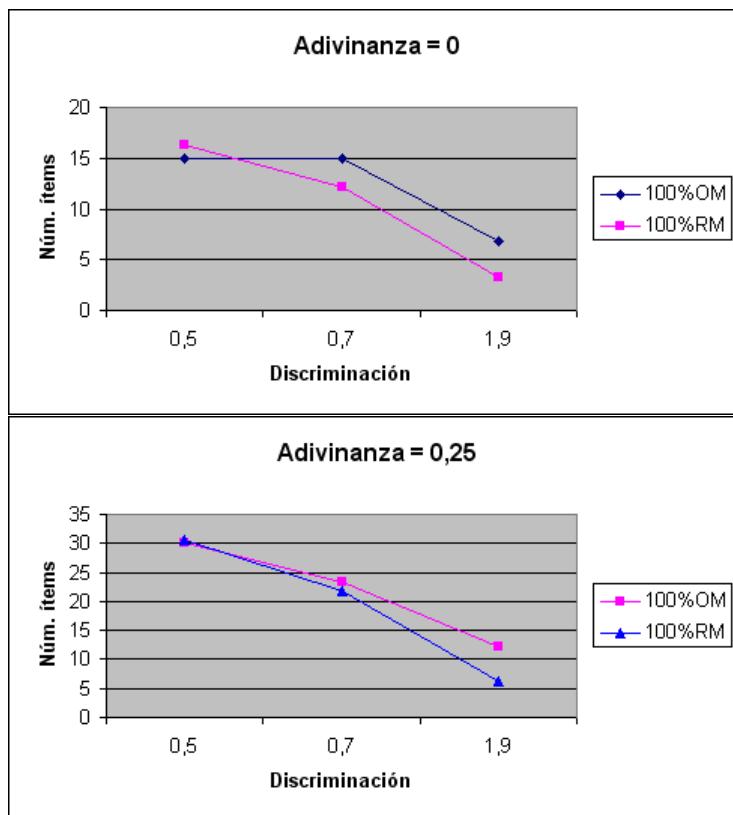


Figura 7.1: Comparación según el número de ítems administrados en el test entre dos bancos de ítems con sólo ítems de opción múltiple o con sólo ítems de respuesta múltiple.

formado únicamente por ítems de respuesta múltiple con opciones independientes. En todos los casos, el banco estaba compuesto por 300 ítems. 100 alumnos simulados fueron sometidos a tests con las características anteriores, en los que además se evaluaba en 12 niveles de conocimiento, utilizando el criterio MAP, y donde el método de finalización era el basado en la máxima precisión esperada con un umbral de 0,001. Se utilizaron tres criterios de selección diferentes: bayesiano, basado en la entropía y basado en la función de información. Las CCO fueron generadas a partir de valores de discriminación obtenidos mediante una función de distribución normal centrada en 1,2, la adivinanza generada a partir de otra centrada en 0,25, y dificultad generada a partir de otra función de distribución normal centrada en el nivel de conocimiento medio.

7.3.4. Resultados

Los resultados de estas simulaciones mostraron que todos los criterios de selección se ven beneficiados cuando en un banco se introducen ítems de respuesta múltiple con opciones independientes. En la tabla 7.2 se muestran, por pares de filas, el número medio de ítems requeridos en las sesiones de evaluación y su intervalo de confianza. Las dos primeras son los resultados cuando se utilizó el criterio de selección bayesiano y las dos segundas cuando

Crit.selecc.		100 % OM	50 % OM	0 % OM
Bayesiano	Núm. ítems	14,46	10,49	8,13
	Error	2,82	0,55	0,38
Entropía	Núm. ítems	11,97	9,12	7,73
	Error	0,64	0,42	0,39
Información	Núm. ítems	33,21	28,01	19,89
	Error	3,17	2,30	0,89

Tabla 7.2: Comparación según el número medio de ítems requerido en una sesión de evaluación, en función del tipo de ítems del banco y del criterio de selección utilizado.

se empleó el basado en la entropía. En las dos terceras se aplicó el método basado en la función de información.

Tal y como se puede ver, las reducciones más significativas cuando se utilizan únicamente ítems de respuesta múltiple con opciones independientes, se producen cuando el criterio de selección es el basado en la entropía esperada. Cuando éste se utiliza, se obtiene una reducción del 64 % de ítems (medio) por sesión. Para los demás, la reducción también es bastante notable: con el mecanismo bayesiano se produce una reducción del 56 %, y con el basado en la función de información una del 60 %.

7.3.5. Discusión

En esta sección se ha puesto de manifiesto que la inclusión de los nuevos tipos de ítems proporcionados por el modelo de respuesta, mejoran significativamente los resultados y el rendimiento de los TAI. Cuando se incluyen ítems de respuesta múltiple con opciones independientes, el número medio de preguntas requerido para diagnosticar el conocimiento del examinando se reduce significativamente. Además, esta reducción no se ve afectada (de forma significativa) por el criterio de selección de ítems que se utilice. En todos ellos, las reducciones que se producen son del orden del 50 % como mínimo.

7.4. Comparación entre los mecanismos de selección de ítems

El objetivo de este estudio es comparar entre sí los mecanismos de selección de ítems incluidos en el modelo de diagnóstico cognitivo. La finalidad es ver, en función de diversos parámetros, cuál de ellos se comporta mejor, en términos de número de ítems requeridos para el diagnóstico y del porcentaje de alumnos correctamente clasificados por el modelo de diagnóstico.

7.4.1. Experimento 1: Comparación entre el criterio bayesiano y el basado en la dificultad

En este experimento, se sometió a una población de 100 alumnos simulados a un test con un banco formado por 300 ítems. Sus CCO fueron construidas de tal forma que se

tuviese un banco lo más rico posible en cuanto a las características psicométricas de sus integrantes. Para ello, la curva característica de la respuesta correcta se generó utilizando el modelo logístico 3PL. El valor del factor de adivinanza era igual a cero para todas las respuestas correctas. El valor de la dificultad se generó aleatoriamente siguiendo una distribución normal igual a la empleada para los niveles de conocimiento de los alumnos. Por último, la discriminación fue generada también siguiendo una distribución normal centrada en 1,2.

Como criterio de finalización del test se utilizó el basado en la máxima precisión esperada. El número de niveles de conocimiento en que se evaluó a los examinandos era igual a doce; y la estimación del conocimiento fue inferida aplicando el criterio MAP.

7.4.2. Resultados

En la tabla 7.3 se muestran los resultados de los experimentos llevados a cabo con ambos criterios de selección de ítems. Se han realizado varias simulaciones en función del valor de umbral de finalización del test (primera columna de la tabla). Las columnas tercera y cuarta contienen el número medio de ítems administrado por individuo; y las quinta y sexta el porcentaje de alumnos clasificados correctamente. Las filas etiquetadas con la palabra *Error* indican el factor de confianza del valor situado en la fila anterior. Como muestran los resultados, el método bayesiano se comporta mejor que el basado en la dificultad en todos los casos en cuanto al número de ítems que requieren para diagnosticar el conocimiento del alumno.

Umbral		Núm. medio ítems		Correc. estim.	
		Bayesiano	Dificultad	Bayesiano	Dificultad
0,1		1,44	3,34	39 %	45 %
	Error	0,21	0,10	7 %	5 %
0,01		4,74	9,09	92 %	92 %
	Error	0,17	0,36	1 %	2 %
0,001		8,79	14,01	99 %	99 %
	Error	0,24	0,63	1 %	0,9 %
0,0001		14,50	18,13	100 %	99 %
	Error	1,10	1,44	0 %	0,3 %
0,00001		20,87	22,51	100 %	100 %
	Error	2,29	1,63	0 %	0 %

Tabla 7.3: Comparación entre los métodos de selección bayesiano (o de la máxima precisión esperada) y el basado en la dificultad.

7.4.3. Experimento 2: Comparación entre el criterio bayesiano y el basado en la entropía

Este experimento es análogo al anterior, con la única salvedad de que se comparan, en este caso, el criterio de selección basado en la máxima precisión esperada con el basado en la entropía.

7.4.4. Resultados

Cada alumno simulado realizó primero un TAI en el que el criterio de selección era el bayesiano de la máxima precisión esperada. Posteriormente, el mismo individuo realizó otro con las mismas características, a excepción del mecanismo de selección utilizado, que en este caso era el basado en la entropía esperada. Para evaluar los resultados de este experimento, se midieron el porcentaje de alumnos correctamente clasificados y el número medio de ítems administrados por examinando.

Umbral		Núm. medio ítems		Correc. estim.	
		Bayesiano	Entropía	Bayesiano	Entropía
0,1		1,44	1,57	39 %	41 %
	Error	0,21	0,10	7 %	6 %
0,01		4,74	4,61	92 %	93 %
	Error	0,17	0,31	1 %	1 %
0,001		8,79	8,34	99 %	99 %
	Error	0,24	0,53	1 %	0,5 %
0,0001		14,50	12,90	100 %	100 %
	Error	1,10	1,45	0 %	0 %
0,00001		20,87	16,59	100 %	100 %
	Error	2,29	1,04	0 %	0 %

Tabla 7.4: Comparación entre los métodos de selección bayesiano de la máxima precisión esperada y basado en la entropía.

La tabla 7.4 muestra la simulaciones llevadas a cabo para diversos niveles de precisión en la estimación. Como se puede ver, cuando el umbral es 0,1, el criterio de selección bayesiano se comporta algo mejor que el basado en la entropía, en cuanto a porcentaje de alumnos correctamente clasificados. Aún así, y especialmente cuanto menor es el umbral de precisión utilizado, se obtienen mejores resultados aplicando el método basado en la entropía. Estas mejoras se ven reflejadas en una reducción del número de ítems medio requerido para diagnosticar el nivel de conocimiento del alumno.

7.4.5. Experimento 3: Comparación entre los criterio bayesiano, basado en la entropía y basado en la información para ítems con diversas propiedades

Los experimentos anteriores han mostrado que los criterios de selección de ítems bayesiano y basado en la entropía ofrecen mejores resultados frente al método basado en la dificultad. En este experimento, basándose en esa premisa, se vuelve a evaluar el comportamiento de esos dos criterios, pero esta vez junto con el basado en la información, en un test en el que el banco de ítems posee cuestiones con características psicométricas más diversas que en el experimento anterior.

Parám.	CCI	Bayesiano				Entropía				Información			
		Dis.	Adi.	N.ítems	Error	Correc.	Error	N.ítems	Error	Correc.	Error	N.ítems	Error
0,2	0	12,73	2,91	99 %	0,7 %	12,29	2,29	99 %	0,4 %	37,39	3,38	99 %	0,8 %
0,2	0,25	23,8	5,47	99 %	0,7 %	21,95	3,64	98 %	1 %	52,04	4,90	99	1 %
0,2	0,5	61,40	8,81	97 %	1 %	53,41	9,79	98 %	1 %	87,85	6,17	97 %	1 %
0,2	0,75	162,67	10,45	87 %	3 %	157,40	9,97	89 %	2 %	174,76	6,49	89 %	2 %
0,2	unif.	36,38	7,30	98 %	1 %	35,05	7,98	97 %	1 %	72,34	7,33	98 %	1 %
0,7	0	13,61	2,95	99 %	0,4 %	13,05	2,94	99 %	1 %	43,29	2,68	98 %	1 %
0,7	0,25	26,23	7,18	98 %	1 %	22,86	4,42	99 %	0,5 %	56,72	6,08	97 %	1 %
0,7	0,5	60,54	9,08	97 %	1 %	62,07	10,27	97 %	1 %	89,77	6,76	97 %	1 %
0,7	0,75	166,08	9,61	91 %	2 %	167,07	4,11	91 %	2 %	180,18	6,66	93 %	1 %
0,7	unif.	39,08	6,29	97 %	1 %	33,00	4,50	98 %	1 %	68,93	7,71	98 %	0,9 %
1,2	0	9,64	0,66	98 %	1 %	8,78	0,58	99 %	0,8 %	27,28	1,66	99 %	0,5 %
1,2	0,25	16,75	4,38	98 %	1 %	16,33	1,55	99 %	1 %	33,50	1,71	99 %	0,5 %
1,2	0,5	36,23	6,79	98 %	1 %	31,31	3,32	98 %	3 %	55,06	4,58	98 %	0,9 %
1,2	0,75	130,18	12,02	94 %	3 %	124,88	9,53	94 %	3 %	146,92	8,77	94 %	2 %
1,2	unif.	19,57	2,28	99 %	1 %	18,81	4,00	98 %	1 %	44,36	3,25	98 %	1 %
1,9	0	6,81	0,32	99 %	1 %	6,38	0,32	99 %	0,6 %	20,55	1,38	99 %	0,5 %
1,9	0,25	12,48	0,70	99 %	0,6 %	11,39	0,36	99 %	0,6 %	22,61	1,90	98 %	0,8 %
1,9	0,5	21,93	2,60	99 %	0,7 %	21,76	2,64	98 %	1 %	36,50	2,44	99 %	0,6 %
1,9	0,75	84,11	9,44	96 %	1 %	83,09	14,66	95 %	2 %	103,34	10,91	96 %	2 %
1,9	unif.	11,70	0,58	98 %	0,9 %	11,15	1,69	99 %	0,8 %	28,37	1,57	99 %	0,7 %

Tabla 7.5: Comparación entre los criterios de selección bayesiano, basado en la entropía y basado en la información, en función de los parámetros de las CCI.

Así, el experimento anterior se lleva a cabo en las mismas condiciones, a excepción de que el umbral de finalización del test no se modifica, tomando siempre el valor 0,001. En este caso, se modifican los parámetros que describen las curvas características de los ítems, menos la dificultad que, al igual que en las pruebas anteriores, se mantiene como un valor generado de forma pseudoaleatoria siguiendo una distribución normal centrada en el nivel de conocimiento medio.

7.4.6. Resultados

La tabla 7.5 muestra los resultados de diversas simulaciones realizadas. Las dos primeras columnas son respectivamente el factor de discriminación y el de adivinanza con el que fueron generadas las CCI de cada ítem. Cuando aparece un valor numérico, éste indica que todas las CCI de los ítems del banco fueron generadas asignando ese valor concreto al parámetro correspondiente. El valor "*unif.*", que aparece en las tres últimas filas de la columna de las adivinanzas, indica que estas últimas en esas simulaciones fueron generadas uniformemente siguiendo una distribución normal centrada en el valor 0,5.

Los resultados muestran también en este caso que, en general, el criterio de selección basado en la entropía se comporta mejor que el bayesiano de la máxima precisión esperada, y que ambos se comportan bastante mejor que el método basado en la información.

En las gráficas de la figura 7.2 se han considerado únicamente los dos criterios de selección con mejores resultados (bayesiano y basado en la entropía). En estas gráficas se ha representado la variación del número medio de ítems requerido en una sesión de evaluación, en función del criterio de adivinanza de los ítems del banco, para los diversos valores de discriminación de éstos. Como se puede apreciar, cuando el valor de la adivinanza es pequeño, la diferencia entre el número de ítems con ambos criterios es prácticamente inapreciable. Por el contrario, en las tres gráficas se muestra que la diferencia es más notable, en favor del criterio basado en la entropía, cuando la adivinanza toma valores mayores.

Las gráficas de las figuras 7.3 y 7.4 son análogas a las anteriores. En este caso, se muestra el número medio de ítems por sesión para ambos criterios de selección, pero esta vez en el eje de abscisas se representa la variación de la discriminación, para una adivinanza determinada. Como se puede apreciar, en general, la diferencia más notable a favor del método de selección basado en la entropía, se manifiesta cuando la discriminación tiene valores medios.

7.4.7. Discusión

A partir de los tres experimentos anteriores se puede inferir que, en general, el criterio de selección basado en la entropía mejora los resultados en lo referente a número de ítems requerido para diagnosticar el nivel de conocimiento del alumno. Conejo et al. (2000) realizaron un estudio empírico similar al segundo experimento que se ha llevado a cabo en esta sección. En este estudio, también se compararon los criterios de selección bayesiano con el método basado en la dificultad para un modelo de respuesta discreto dicotómico. Al igual que el estudio realizado, este análisis concluyó que el criterio bayesiano se comporta mejor que el basado en la dificultad, tanto en el número de ítems requerido como en porcentaje de alumnos correctamente diagnosticados. Aún así, también es cierto que mientras que el criterio de selección bayesiano es más costoso computacionalmente, el número de operaciones requerida por el criterio basado en la dificultad es sustancialmente inferior.

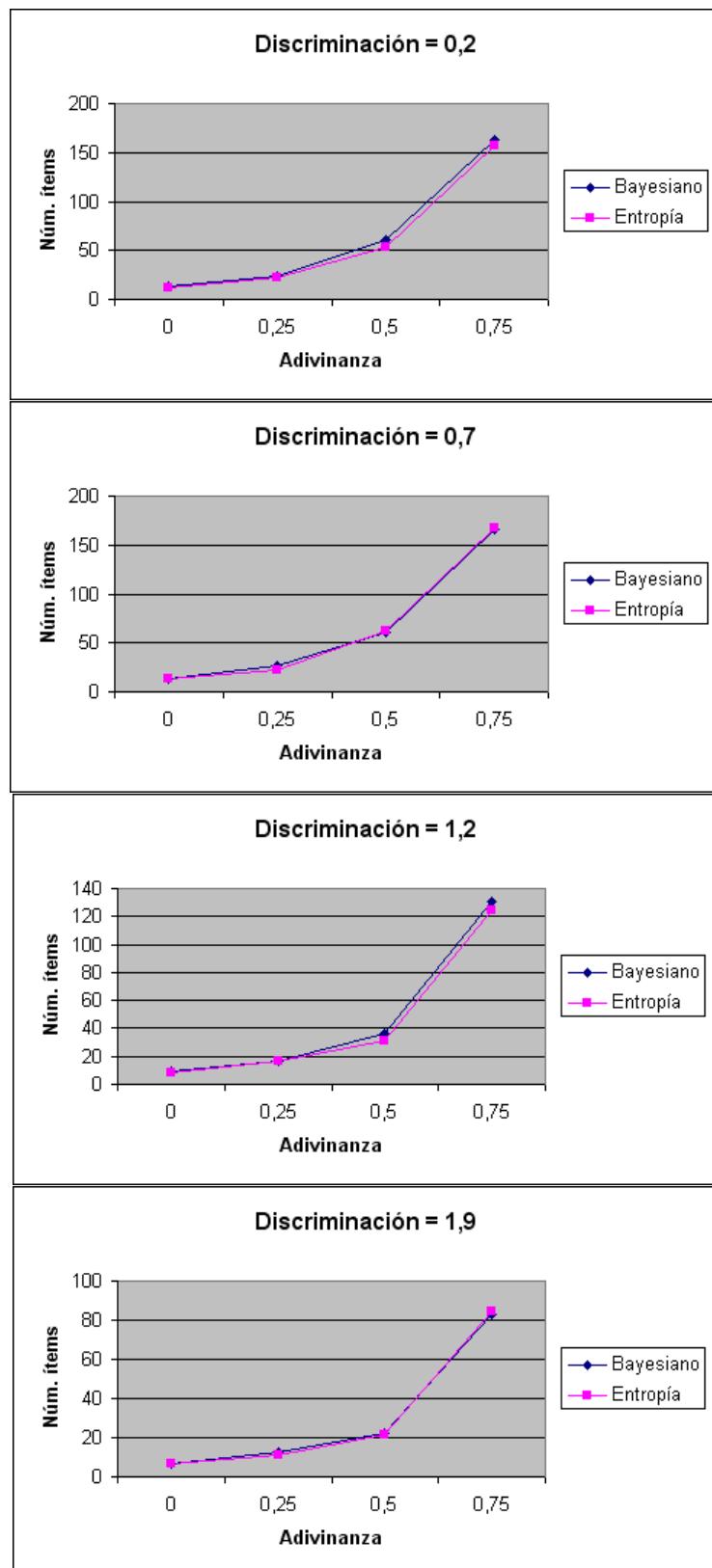


Figura 7.2: Comparación en función del número de ítems del test entre los criterios de selección bayesiano y basado en la entropía para una discriminación determinada en función de la adivinanza.

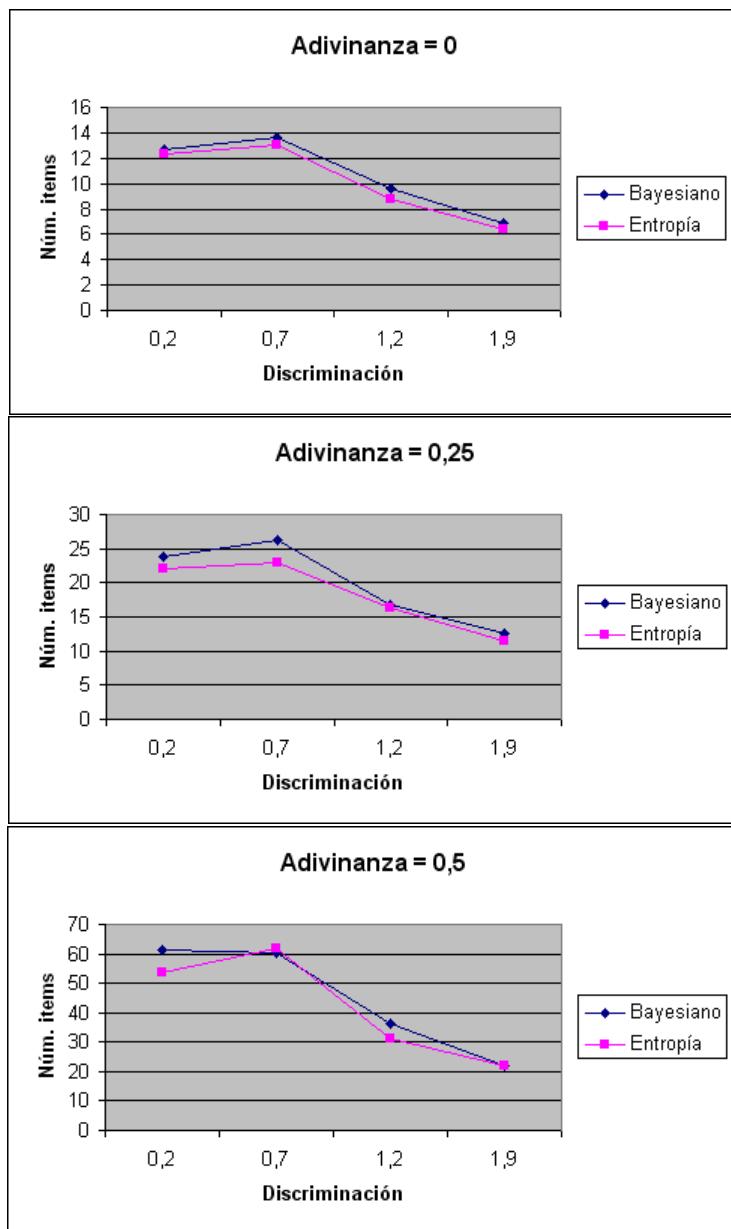


Figura 7.3: Comparación en función del número de ítems del test, entre los criterios de selección bayesiano y basado en la entropía, para una adivinanza determinada, en función de la discriminación (I parte).

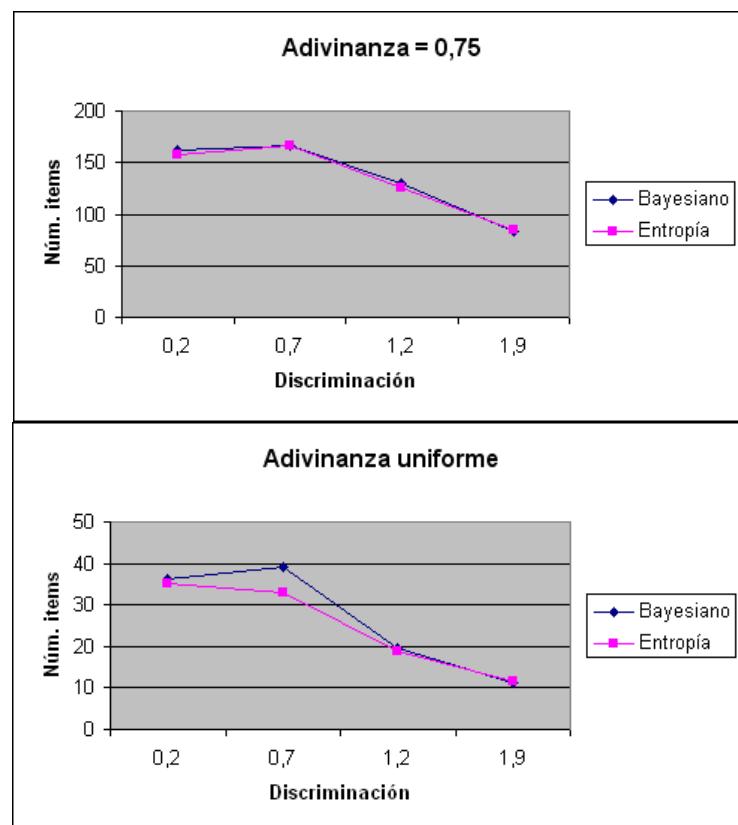


Figura 7.4: Comparación en función del número de ítems del test, entre los criterios de selección bayesiano y basado en la entropía, para una adivinanza determinada, en función de la discriminación (II parte).

7.5. Comparación entre los modelos de respuesta propuesto y el clásico 3PL continuo

El objetivo de este experimento es comparar el modelo TRI más comúnmente utilizado, esto es, el dicotómico basado en la función 3PL, con el propuesto en esta tesis. Para ello, se han realizado diversas simulaciones en las que se ha estudiado cómo se comporta un modelo frente al otro, en función de dos factores: el número medio de ítems requerido para el diagnóstico del conocimiento de los alumnos; y posteriormente, del tiempo de cómputo requerido para seleccionar el ítem, actualizar el conocimiento del alumno tras su respuesta, y determinar si el test debe finalizar.

7.5.1. Eficiencia según el número de ítems

En este caso, se comparan ambos modelos teniendo en cuenta el número de ítems requeridos en una sesión de evaluación para inferir el nivel conocimiento del examinando con un precisión determinada a priori.

Experimento 1: Comparación entre los modelos de respuesta dicotomizado y el clásico 3PL

En este análisis se ha utilizado una población de 100 examinandos simulados para llevar a cabo el estudio. Cada alumno disponía de un banco de 300 ítems. El objetivo era medir su nivel de conocimiento en un test sobre un único concepto. Como es habitual, cada alumno fue generado con un nivel de conocimiento real a priori, que determinó su comportamiento durante la realización del test. Asimismo se asumía que el banco estaba calibrado correctamente. Los parámetros de sus ítems fueron generados de forma pseudoaleatoria siguiendo una distribución normal, que aseguraba un amplio abanico de diversas dificultades y discriminaciones. Todos los ítems generados eran de opción múltiple con tres posibles respuestas, factor de adivinanza igual a cero, y con discriminación basada en una distribución normal centrada en 1, 2. Además en ellos se permitía al alumno dejar la respuesta en blanco.

Se definieron dos tests bastante similares. Ambos fueron configurados con el criterio de selección de ítems bayesiano. El método de finalización era el basado en la precisión de la estimación. La inicialización del modelo del alumno era siempre una distribución equiprobable. En cuanto a las diferencias entre los dos tests, éstas consistían en lo siguiente: el primero de ellos llevaba a cabo la evaluación propuesta en el modelo de esta tesis, con la particularidad de que en vez de utilizarse el modelo de respuesta de forma polítómica, se empleaba una versión dicotómica del mismo. Para dicotomizarlo, bastó con asignar a las opciones de respuesta de cada ítem, diferentes de la opción correcta, una CCO igual a la opuesta de la curva de la respuesta correcta. Para asegurar que cada ítem fuese un espacio probabilístico, las probabilidades de la CCO de cada respuesta incorrecta fueron divididas por el número de respuestas incorrectas. De esta forma se aseguraba que la suma de todas las curvas del ítem era igual a uno.

El segundo test utilizaba los mismos criterios de selección de ítems, finalización y evaluación que el anterior, pero en este caso éstos eran continuos y dicotómicos. El nivel de conocimiento para este caso oscilaba entre -4 y 4, y también fue generado siguiendo una distribución normal.

Resultados

Los resultados del estudio se muestran en la tabla 7.6. Ésta muestra diversas simulaciones realizadas, modificando el valor del umbral de precisión de la estimación, que determina cuando puede finalizar el test. En la tabla se muestran el número medio de ítems administrado a cada alumno para los dos tests, y el porcentaje de aciertos en la estimación (en comparación con el nivel de conocimiento con el que el examinando fue generado). Junto con ambos datos se incluye también el correspondiente factor de confianza. Como se puede apreciar, cuanto menor es el valor del umbral, el número de ítems que deben suministrarse al alumno, para estimar su conocimiento, es mayor.

Umbral	Discreto				Continuo			
	N.ítems	Error	Correc.	Error	N.ítems	Error	Correc.	Error
0,25	1,51	0,05	98,00 %	0,06 %	1,00	0,00	96,91 %	0,04 %
0,2	2,00	0,00	98,81 %	0,04 %	2,00	0,00	98,73 %	0,07 %
0,15	2,00	0,00	98,79 %	0,05 %	2,33	0,06	98,80 %	0,07 %
0,1	2,00	0,00	98,77 %	0,05 %	2,43	0,05	98,84 %	0,06 %
0,01	7,78	0,19	98,88 %	0,04 %	6,06	0,12	99,04 %	0,05 %
0,001	18,76	2,04	99,99 %	0,00 %	9,53	0,25	99,05 %	0,04 %
0,0001	37,68	1,55	99,99 %	0,00 %	12,99	0,26	98,99 %	0,03 %
0,00001	51,03	10,79	100,00 %	0,00 %	17,64	2,95	99,06 %	0,09 %

Tabla 7.6: Comparación entre los modelos clásico 3PL y con el propuesto en esta tesis en su versión dicotómica.

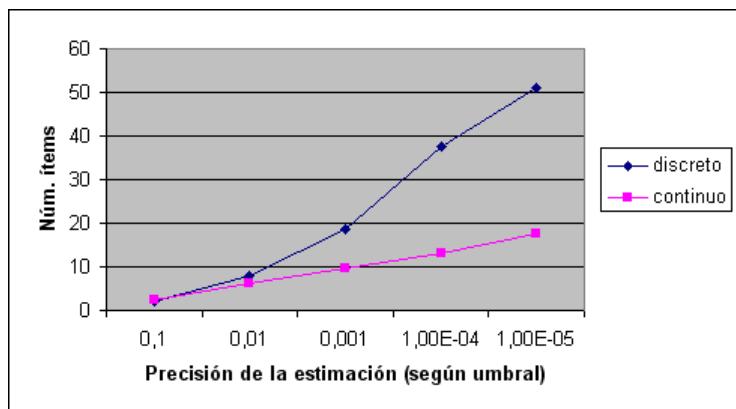


Figura 7.5: Comparación según el número de ítems requerido para evaluar entre los modelos 3PL continuo y el propuesto en esta tesis dicotomizado.

En la figura 7.5 se han representado, de forma comparativa, el número medio de ítems requerido por ambos modelos para diagnosticar el conocimiento del alumno, según el umbral indicado en el eje de abscisas. Obsérvese que los valores de ese eje están en escala logarítmica inversa, por lo que son menores conforme se avanza hacia la derecha. El motivo de esto es que lo que se ha intentado representar de forma creciente es la precisión requerida en el diagnóstico. La gráfica de color rosa representa los valores para el modelo continuo 3PL,

mientras que la azul los correspondientes a la versión dicotomizada del modelo presentado en esta tesis.

Los datos evidencian que, para niveles de precisión superiores a la centésima, los resultados son similares en todos los aspectos: número de ítems, precisión, etc. Por el contrario, para valores inferiores, se puede apreciar que la discretización obliga a administrar un número desmesurado de ítems para obtener una estimación con la precisión requerida.

Experimento 2: Comparación entre el modelo de respuesta (politómico) y el modelo clásico 3PL

Tras los resultados obtenidos en el experimento anterior, se realizó este segundo experimento en el que se comparaba el modelo continuo anterior con el discreto, pero esta vez, en su versión politómica. Según esto, a excepción de la salvedad mencionada, las condiciones del experimento eran exactamente las mismas.

Resultados

La tabla 7.7 muestra los datos obtenidos tras el experimento. Como se puede apreciar, el modelo en su versión politómica ofrece unos resultados bastante similares al modelo continuo. El número medio de ítems necesario para evaluar al alumno según el nivel de precisión indicado, se ha reducido considerable con respecto al modelo en versión dicotómica. Incluso para algunos casos se reduce el número de ítems, manteniéndose o mejorándose la cantidad de individuos correctamente evaluados.

Umbral	Discreto				Continuo			
	N.ítems	Error	Correc.	Error	N.ítems	Error	Correc.	Error
0,25	1,05	0,15	98,31 %	0,29 %	1,00	0,00	96,94 %	0,03 %
0,2	1,24	0,20	98,51 %	0,17 %	2,00	0,00	98,69 %	0,06 %
0,15	1,63	0,14	98,98 %	0,18 %	2,33	0,02	98,82 %	0,07 %
0,1	1,73	0,22	99,13 %	0,20 %	2,40	0,05	98,79 %	0,06 %
0,01	5,74	0,20	99,92 %	0,03 %	6,08	0,08	99,08 %	0,03 %
0,001	9,20	0,10	99,99 %	0,01 %	9,37	0,24	99,04 %	0,06 %
0,0001	15,41	2,03	100,00 %	0,00 %	13,05	0,26	99,98 %	0,03 %
0,00001	21,01	5,44	100,00 %	0,00 %	17,29	3,04	99,45 %	0,03 %

Tabla 7.7: Comparación entre los modelos clásico 3PL y el propuesto en esta tesis.

La figura 7.6 muestra la evolución, en función de la precisión requerida para el diagnóstico, del número medio de ítems empleado en cada sesión de evaluación. Al igual que en la figura 7.5, el eje de abscisas está en escala logarítmica inversa. Como se puede ver, cuando para requisitos de precisión en las estimaciones son menores, ambos modelos hacen uso de un número medio de ítems similar. Por el contrario, cuando los requisitos son mayores, la cantidad de ítems que necesita el modelo presentado en esta tesis se dispara frente al valor que se precisa con el modelo 3PL.

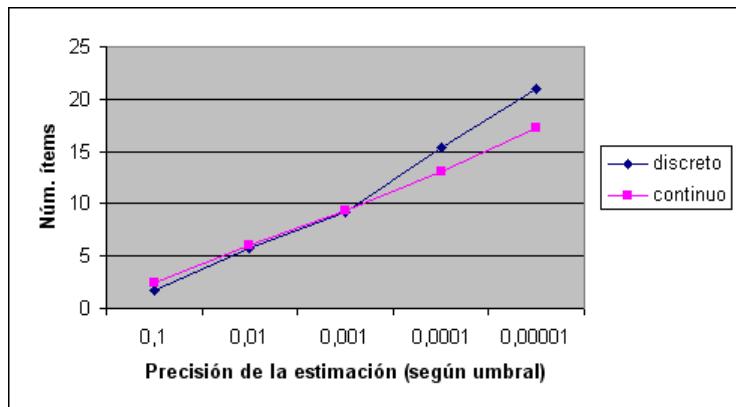


Figura 7.6: Comparación según el número medio de ítems requerido para evaluar entre los modelos 3PL continuo y el propuesto en esta tesis.

Experimento 3: Comparación entre los modelos de respuesta (polítómico) con selección basada en la entropía y el clásico 3PL

Este experimento es análogo al anterior, la única modificación que se ha hecho es que, cuando se utiliza el modelo presentado en esta tesis, el criterio de selección de ítems aplicado es el basado en la entropía. El modelo continuo basado en la 3PL sigue utilizando el método bayesiano.

Resultados

Umbral	Discreto				Continuo			
	N.ítems	Error	Correc.	Error	N.ítems	Error	Correc.	Error
0,25	1,03	0,14	98,20 %	0,20 %	1,00	0,00	96,93 %	0,02 %
0,2	1,19	0,22	98,49 %	0,36 %	2,00	0,00	98,69 %	0,03 %
0,15	1,56	0,20	98,83 %	0,31 %	2,33	0,04	98,84 %	0,07 %
0,1	1,72	0,19	99,09 %	0,15 %	2,43	0,07	98,82 %	0,07 %
0,01	5,74	0,15	99,91 %	0,03 %	6,04	0,16	99,02 %	0,05 %
0,001	8,83	1,02	99,99 %	0,01 %	9,37	0,27	99,07 %	0,03 %
0,0001	11,75	1,19	100,00 %	0,00 %	13,05	0,26	99,98 %	0,03 %
0,00001	15,03	2,90	100,00 %	0,00 %	17,29	3,04	99,45 %	0,03 %

Tabla 7.8: Comparación entre los modelos clásico 3PL y el propuesto en la tesis, utilizando el criterio de selección basado en la entropía.

La tabla 7.8 muestra los resultados obtenidos en este experimento. Como se puede ver, en este caso los resultados son ligeramente mejores para el modelo presentado en esta tesis en comparación con el discreto basado en la 3PL.

La figura 7.7 es análoga a la 7.6, con la única excepción de que el método de elección de ítems utilizado cuando se aplica el modelo de diagnóstico discreto de esta tesis es el basado

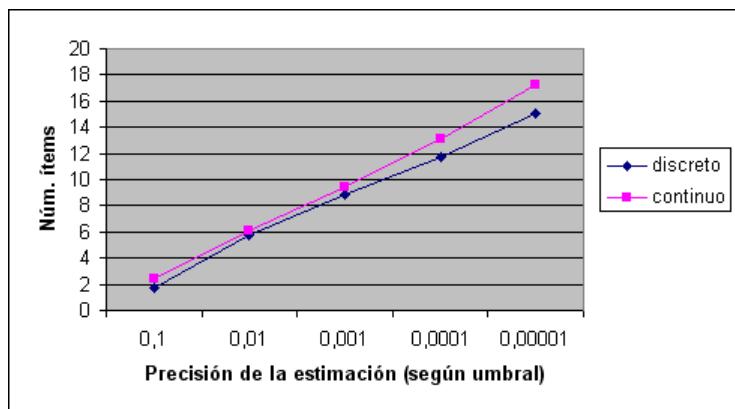


Figura 7.7: Comparación según el número medio de ítems requerido para evaluar entre el modelo 3PL continuo y el propuesto en esta tesis, con criterio de selección basado en la entropía.

en la entropía. Como puede verse, este criterio mejora los resultados del TAI con respecto a los que se obtienen con el modelo 3PL continuo, incluso para los valores de precisión mayores.

7.5.2. Eficiencia según el tiempo de cómputo empleado

Una segunda comparación entre los modelos clásico 3PL continuo y el propuesto en esta tesis, se realizará en virtud del tiempo de cómputo empleado en las fases principales del algoritmo que describe el funcionamiento de un TAI; es decir, selección del ítem, actualización del conocimiento del alumno y decisión de finalización del test.

Conseguir tiempos reducidos en la ejecución de estas fases es fundamental, como lo es en cualquier sistema informático que opere en tiempo real. En las aplicaciones ubicadas en la Web la importancia del tiempo de cómputo es si cabe aún mayor, ya que a éste hay que sumarle el empleado en las comunicaciones a través de la red.

Resultados

La tabla 7.9 muestra los tiempos medios obtenidos en cada una de las tres fases principales del algoritmo TAI, junto con su factor de confianza. Para tomar esos datos, se llevó a cabo un test simulado sobre una muestra poblacional de 100 individuos y basado en un banco de 300 ítems. Éstos eran de opción múltiple con tres posibles respuesta. Los parámetros de sus CCI fueron generados siguiendo distribuciones normales centradas en: 1, 2, la discriminación; el número medio de la escala de niveles de conocimiento, la dificultad; y 0,25, la adivinanza.

Se realizaron diversas simulaciones variando el número de niveles de conocimiento en los que los alumnos eran evaluados (para el caso discreto), y modificando también el umbral que determina la finalización del test. El criterio de selección empleado era el bayesiano, el de inferencia del nivel de conocimiento el MAP, y por último, la finalización se basaba en la máxima precisión esperada.

umbral	núm. niveles	Continuo			Discreto			
		selección	actualiz.	finaliz.	selección	actualiz.	finaliz.	
0,1	2	tiempo	380,95	0,21	0,62	4,53	0,01	0,08
		error	8,63	0,03	0,05	0,09	0,009	0,02
0,1	3	tiempo	378,79	0,21	0,60	4,62	0,02	0,08
		error	7,72	0,03	0,05	0,09	0,002	0,05
0,1	6	tiempo	380,91	0,23	0,66	4,92	0,02	0,09
		error	10,24	0,006	0,008	0,10	0,002	0,007
0,1	12	tiempo	369,01	0,22	0,65	5,43	0,02	0,11
		error	8,78	0,004	0,008	0,001	0,002	0,002
0,1	24	tiempo	381,27	0,20	0,63	6,46	0,02	0,12
		error	6,07	0,003	0,003	0,36	0,001	0,009
0,1	48	tiempo	379,44	0,20	0,64	8,35	0,02	0,19
		error	4,29	0,002	0,005	0,70	0,0009	0,02
0,1	100	tiempo	382,16	0,19	0,65	12,20	0,02	0,25
		error	3,27	0,001	0,003	1,47	0,001	0,01
0,01	2	tiempo	386,07	0,20	0,61	4,47	0,02	0,13
		error	4,59	0,02	0,06	0,20	0,01	0,03
0,01	3	tiempo	398,18	0,23	0,65	4,54	0,02	0,14
		error	4,15	0,002	0,01	0,03	0,002	0,007
0,01	6	tiempo	394,44	0,23	0,68	4,75	0,02	0,14
		error	5,46	0,001	0,02	0,07	0,002	0,01
0,01	12	tiempo	389,37	0,24	0,77	5,21	0,02	0,25
		error	0,68	0,003	0,04	0,15	0,0007	0,05
0,01	24	tiempo	382,42	0,24	0,93	5,90	0,02	0,38
		error	0,60	0,002	0,03	0,23	0,0009	0,02
0,01	48	tiempo	388,93	0,25	1,06	7,19	0,02	0,44
		error	5,06	0,005	0,05	0,51	0,0004	0,02
0,01	100	tiempo	371,71	0,27	1,34	9,95	0,02	0,58
		error	6,66	0,004	0,09	1,06	0,002	0,04
0,001	2	tiempo	371,71	0,2	0,88	4,73	0,04	0,17
		error	4,59	0,02	0,07	0,07	0,007	0,01
0,001	3	tiempo	386,51	0,21	0,78	4,80	0,03	0,18
		error	4,91	0,01	0,04	0,05	0,004	0,006
0,001	6	tiempo	379,02	0,21	0,69	5,14	0,02	0,19
		error	12,33	0,007	0,01	0,11	0,001	0,008
0,001	12	tiempo	380,23	0,20	0,63	5,69	0,2	0,19
		error	9,39	0,002	0,01	0,17	0,001	0,008
0,001	24	tiempo	382,89	0,19	0,61	6,62	0,02	0,27
		error	6,2	0,001	0,009	0,34	0,001	0,02
0,001	48	tiempo	377,18	0,19	0,60	8,20	0,023	0,31
		error	4,48	0,0009	0,006	0,001	0,001	0,01
0,001	100	tiempo	383,76	0,19	0,61	13,13	0,024	0,38
		error	3,08	0,001	0,007	1,58	0,0009	0,02

Tabla 7.9: Tiempo de cómputo de las fases del algoritmo TAI expresado en milisegundos.

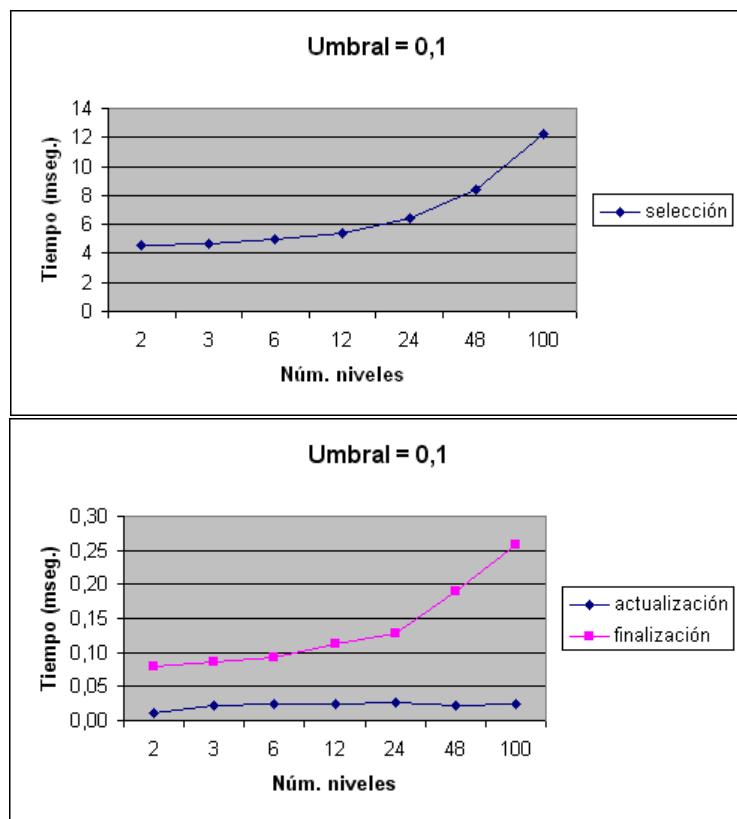


Figura 7.8: Tiempo empleado por el modelo propuesto en cada fase del algoritmo adaptativo, cuando el umbral es 0,1.

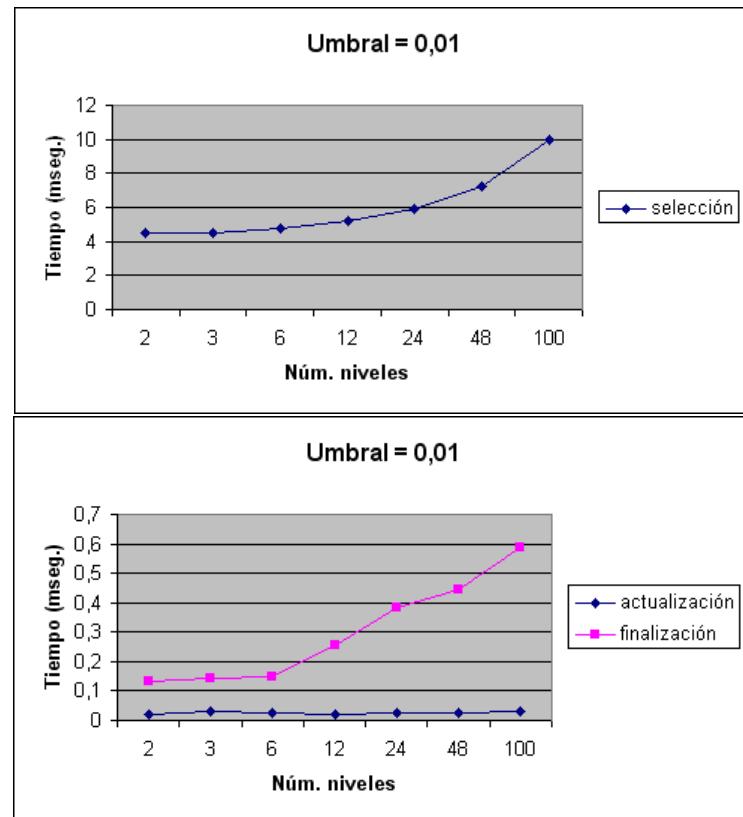


Figura 7.9: Tiempo empleado por el modelo propuesto en cada fase del algoritmo adaptativo, cuando el umbral es 0,01.

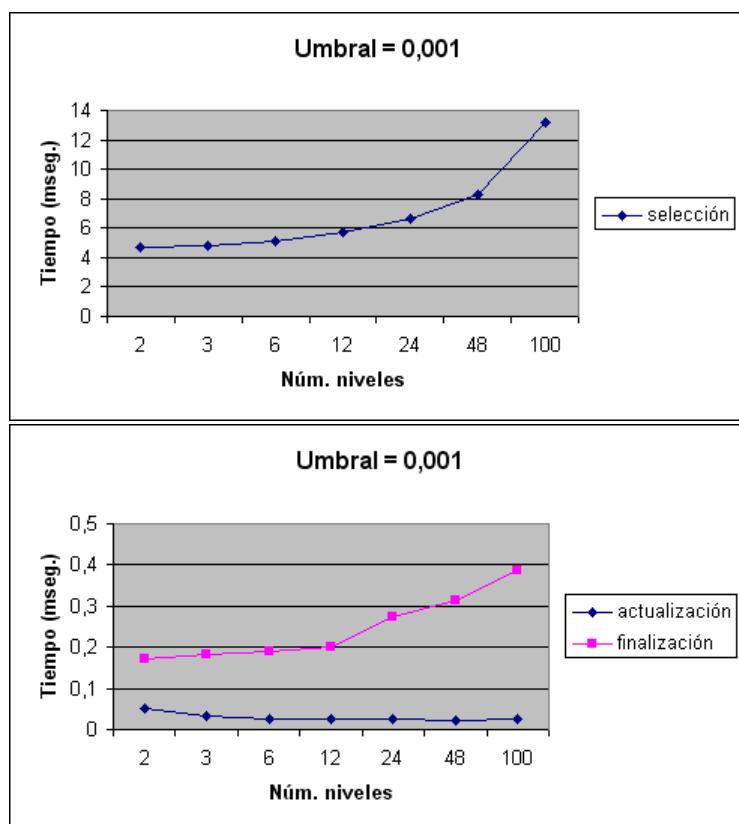


Figura 7.10: Tiempo empleado por el modelo propuesto en cada fase del algoritmo adaptativo, cuando el umbral es 0,001.

Como se puede ver, el tiempo necesario para determinar el ítem que debe ser mostrado al alumno es siempre considerablemente mayor en el caso continuo frente al discreto. De hecho, la diferencia se traduce en que el algoritmo continuo necesita del orden de cien veces más tiempo de cómputo que el discreto. Por otro lado, las fases de actualización de la estimación temporal del conocimiento del alumno, y de comprobación de si el test debe finalizar también requieren menor tiempo de cómputo aunque, en ambas las diferencias no son tan significativas.

En las figuras 7.8, 7.9 y 7.10 se muestra la evolución de los tiempos, en función del número de niveles de conocimiento empleados. En ellas se observa que, aunque los valores se ven incrementados conforme aumenta el número de niveles. A pesar de ello, incluso para un total de 100 niveles, los tiempos son aceptables, especialmente cuando se comparan con los obtenidos para la aproximación continua.

7.5.3. Discusión

En este estudio se ha comparado el modelo de diagnóstico presentado en esta tesis con el clásico basado en la función 3PL. Inicialmente, esta tarea se ha llevado a cabo en términos del número medio de ítems requerido por sesión de evaluación. Como se ha puesto de manifiesto, el modelo presentado en esta tesis mejora los resultados que se obtienen con el 3PL, aún siendo discreto. Esta mejora sólo se produce cuando el criterio de selección que se utiliza es el basado en la entropía. Por el contrario, si el método de elección es el bayesiano, aunque para precisiones menores los resultados son similares; cuando los requisitos de precisión son elevados, el modelo 3PL es considerablemente mejor.

Por otra parte, se ha estudiado el tiempo empleado en cada fase del algoritmo de aplicación de un TAI, utilizando sendos modelos. Como resultado, se ha evidenciado que el modelo propuesto es practicable desde el punto de vista de sus requisitos en tiempo real; algo que es fundamental en especial cuando se construye una herramienta cuya plataforma de ubicación es la Web. La fase del algoritmo TAI que más tiempo consume, la selección del ítem, en el modelo discreto y polítómico propuesto requiere un tiempo de cómputo del orden de decenas de milisegundos, mientras que para el modelo 3PL continuo y dicotómico éste se ve incrementado hasta el orden de las centenas.

7.6. Evaluación simultánea de múltiples conceptos en un mismo test

El objetivo de este estudio es demostrar, de forma empírica, que los criterios de selección del modelo de diagnóstico propuesto llevan a cabo una elección balanceada de ítems en función de los conceptos que evalúan. Por consiguiente, la finalidad es mostrar que cuando en un mismo test se persigue diagnosticar el conocimiento en más de un concepto de forma simultánea, el criterio de selección por sí solo es capaz de ir eligiendo ítems de diversos conceptos, de forma que, al final del test, el número de los que se hayan mostrado de cada concepto, sea el mínimo necesario para realizar una estimación precisa. En el criterio basado en la dificultad y en el basado en la información no es necesario demostrarlo ya que, por propia definición (como se vio en la sección 5.3.2), se selecciona el ítem de aquel concepto cuya distribución sea más imprecisa. Por este motivo, este experimento se centra únicamente en los criterios bayesiano y basado en la entropía.

7.6.1. Experimento

En este experimento se realizaron diversas pruebas en las que se sometió a un conjunto de alumnos simulados a un test en el que se evaluaban varios conceptos de forma simultánea. Por cada examinando, se analizó, en cada momento, qué concepto debía ser evaluado, en función de la precisión de su distribución probabilística de conocimiento estimada. De esta forma, antes de proceder a seleccionar un ítem, se comprobó cuál era la precisión de la curva probabilística del conocimiento del alumno en los conceptos evaluados en el test. Se comprobó además cuál era el concepto cuya estimación era menos precisa. A continuación, se llevó a cabo la selección siguiendo el criterio bayesiano o el basado en la entropía, según correspondiera. Posteriormente, se comprobó si el ítem elegido evaluaba al concepto cuya estimación era más precisa. En los casos en los que la predicción era incorrecta, se calculó la diferencia entre la precisión de la estimación del concepto que debía haber sido evaluado según la predicción, y la precisión a priori (antes de calcular la nueva estimación tras administrar al alumno el nuevo ítem seleccionado) del concepto que evalúa el ítem seleccionado.

Los alumnos fueron evaluados en doce niveles de conocimiento. Los bancos empleados en cada simulación contenían ítems de opción múltiple, y fueron creados, a partir de la función 3PL. Los valores de los parámetros de la CCI fueron asignados de la siguiente forma: el factor de discriminación fue generado siguiendo una distribución normal centrada en 1, 2; la dificultad siguiendo una uniforme centrada en 5; por último, el factor de adivinanza siguiendo una distribución normal centrada en 0, 25.

7.6.2. Resultados

selec.	núm. concep.	núm. ítems	ítems incorr. sel.	dif. disper.	núm. ítems	correc. estim.
bayes.	3	3000	4,54 %($\pm 1,12$)	0,0003079	9,99($\pm 2,03$)	98,79 %
bayes.	5	5000	4,83 %($\pm 1,05$)	0,0002431	10,73($\pm 2,66$)	99,20 %
bayes.	7	7000	5,21 %($\pm 0,99$)	0,0001999	10,53($\pm 2,87$)	99,28 %
bayes.	9	9000	5,42 %($\pm 1,22$)	0,0001986	10,20($\pm 1,99$)	99,43 %
entrop.	3	3000	29,55 %($\pm 4,59$)	0,0163988	8,99($\pm 2,13$)	98,66 %
entrop.	5	5000	33,61 %($\pm 5,27$)	0,0181101	9,25($\pm 3,42$)	99,80 %
entrop.	7	7000	32,97 %($\pm 3,98$)	0,0142118	9,42($\pm 2,86$)	99,57 %
entrop.	9	9000	34,22 %($\pm 6,10$)	0,0151362	9,87($\pm 2,44$)	99,91 %

Tabla 7.10: Resultados de las simulaciones de tests sobre múltiples conceptos.

La tabla 7.10 muestra los resultados de los diferentes experimentos. Cada fila representa una simulación llevada a cabo con un test de diferentes características. Todos los tests fueron realizados por 100 alumnos simulados. La primera columna de la tabla indica el criterio de selección utilizado, esto es, bayesiano o basado en la entropía. En la segunda se indica el número de conceptos evaluados de forma simultánea. Se han realizado tests con 3, 5, 7 y 9 conceptos respectivamente. La tercera columna recoge el número total de ítems del banco utilizado en cada test. Cuando el banco fue creado, sus cuestiones fueron siendo asignadas consecutivamente a cada uno de los conceptos del test, de forma que todos estos tuviesen asignados directamente el mismo número de ítems, concretamente 1000. La

calificación de cada examinando fue obtenida aplicando el criterio MAP, y el método de finalización empleado el basado en la máxima precisión esperada con un umbral de 0,001.

La cuarta columna de la tabla 7.10 indica el porcentaje medio de veces que no se seleccionó un ítem que evaluaba al concepto en el que la estimación del conocimiento era menos precisa, junto con su intervalo de confianza entre paréntesis. Asimismo, la quinta muestra la diferencia media entre las precisiones de la distribución del conocimiento en el concepto con menor valor, y la correspondiente al que evaluaba el ítem seleccionado. La sexta columna contiene el número medio de cuestiones por concepto administrados al alumno en el test, junto con su factor de confianza. Por último, la séptima expresa el porcentaje medio de alumnos correctamente clasificados por concepto al final del test, es decir, que su conocimiento real sobre el concepto coincidía con el nivel inferido por el TAI.

7.6.3. Discusión

Como se puede apreciar en la tabla 7.10, para ambos criterios de selección, el número de ítems necesario para estimar el conocimiento del alumno por concepto es bastante similar. La principal diferencia en la aplicación de ambos criterios reside en cómo se lleva a cabo la elección de la siguiente cuestión durante la realización del test. Para las simulaciones que utilizan el criterio bayesiano, el siguiente ítem seleccionado, en un mínimo del 95 % de las veces, estaba asociado al concepto cuya distribución estimada del conocimiento del alumno era la menos precisa. En el 5 % restante éste estaba asociado a otro cuya distribución de conocimiento poseía una precisión pequeña, la cual además era bastante similar en valor a la de menor precisión. La quinta columna muestra que la diferencia media entre dispersiones es bastante pequeña en los casos en los que la elección no es la esperada (nunca superior a una diezmillésima).

Por otra parte, en las simulaciones en las que se ha utilizado el criterio basado en la entropía, el porcentaje de veces que se selecciona un ítem asociado al concepto cuya estimación tiene menor precisión es menor (en torno al 65 % de las veces). En estas ocasiones se selecciona un ítem que está asociado al concepto cuya distribución tiene menor precisión o bien al siguiente con menor valor de precisión, siendo la diferencia entre éstas del orden de una centésima. Aún así, y tal y como se puede ver en la tabla de resultados, el número de ítems requerido por ambos criterios es bastante similar, obteniéndose en todos los casos estimaciones considerablemente certeras del conocimiento de los alumnos en los conceptos.

Estos resultados ponen de manifiesto que los criterios de selección bayesiano y basado en la entropía son capaces, por sí solos de realizar una selección balanceada en contenido de los ítems. En estos tests la selección se lleva a cabo de forma balanceada sin necesidad de recurrir a heurísticos.

7.7. Estudio del método de calibración de ítems

El objetivo de esta sección es estudiar el comportamiento del método de calibración propuesto para los ítems del modelo de respuesta. Se estudiará la precisión de las estimaciones de las curvas características a partir de diversos conjuntos de sesiones de tests. Asimismo, y como efecto colateral, se intenta mostrar que, gracias a este algoritmo de calibración, el modelo presentado es factible. Para ello, bastará con demostrar que con una muestra de alumnos inicial con número razonable de individuos los resultados de la calibración son satisfactorios.

7.7.1. Experimento 1: Estudio del valor adecuado para el parámetro de suavizado y de la bondad del método de calibración

El objetivo de este experimento es estudiar el rendimiento del algoritmo de calibración. Se estudiará como afecta el valor del parámetro de suavizado a los resultados de la calibración. Con este fin, se partirá de un conjunto de ítems cuyas CCO se conocen a priori, y a partir de diversas muestras de alumnos simulados, se procederá a calibrar esas CCO. Finalmente, se calculará el ECM entre las CCO reales y las calibradas como indicador del error cometido en la estimación, aplicando la siguiente fórmula:

$$ECM_{ij} = \sqrt{\sum_{k=0}^{K-1} \frac{[P_{ij}(S(r_j) = 1|k) - P_{ij}^R(S(r_j) = 1|k)]^2}{K}} \quad (7.2)$$

donde $P_{ij}^R(S(r_j) = 1|k)$ es el valor de la CCO real correspondiente a la opción j-ésima del ítem i-ésimo, en el nivel k . El número total de niveles de conocimiento es K .

Además, dentro del proceso de calibración, se inferirá también el nivel de conocimiento de los alumnos pertenecientes a la muestra tomada como punto de partida para realizar esta tarea. El valor inferido será comparado con el nivel de conocimiento real con el que el alumno simulado fue generado. Asimismo, se procederá, a partir de las CCO calibradas, a administrar un TAI a una nueva muestra de individuos, y se estudiará el porcentaje de ellos cuyo nivel de conocimiento se diagnostica de forma correcta.

A los alumnos de la muestra utilizada se les administra un test heurístico, formado por los 50 ítems que se desea calibrar. Éstos son la mitad de opción múltiple con tres opciones de respuesta, y la otra parte de respuesta múltiple con cuatro opciones de respuesta independientes. El objetivo es, por tanto, inferir un total de 200 CCO. El número de niveles de conocimiento utilizados como rango de éstas, y como escala de evaluación es igual a seis. Los parámetros de las curvas reales han sido generados, siguiendo el procedimiento indicado en la sección 7.2. Los valores del factor de discriminación, dificultad y adivinanza han sido generados pseudoaleatoriamente siguiendo distribuciones normales centradas en 1, 2, 2 y 0,15, respectivamente.

Tras aplicar el proceso de calibración, se procede a administrar un TAI, en el que el banco de ítems está formado única y exclusivamente por los 50 calibrados en la etapa anterior. El criterio de selección de ítems de este test es el basado en la entropía, y el método de finalización el de la máxima precisión esperada, al que se aplica un umbral de 0,001. Por último, el mecanismo de evaluación empleado es el MAP. El TAI con estas características se administra a una nueva muestra de 100 alumnos simulados. En este caso, los examinandos también son evaluados en seis niveles de conocimiento.

7.7.2. Resultados

La tablas de la 7.11 a la 7.16 recogen los resultados de las calibraciones realizadas con muestras de alumnos de diferentes tamaños: 20, 50, 100, 300, 500, 1000, 5000 y 10000 individuos, tal y como se muestra en la primera columna. La segunda representa el valor del parámetro de suavizado utilizado en cada proceso de calibración. Cada fila contiene, de esta forma, los resultados de la aplicación del algoritmo de calibración bajo condiciones

diferentes. Todas aquellas filas en las que el número de alumnos es el mismo, representan simulaciones del proceso de calibración llevadas a cabo utilizando la misma muestra poblacional de examinandos.

Como se puede apreciar, el rango de valores del parámetro de suavizado representados en la tabla depende del tamaño de la muestra de alumnos. Tras multitud de simulaciones, se ha determinado el rango de valores de este parámetro en el que se llega a un mínimo, entendido éste como el valor en el que los resultados de la calibración son los mejores, para el tamaño muestral correspondiente.

La tercera columna de la tabla presenta el número de alumnos de la muestra cuyo nivel de conocimiento ha sido inferido correctamente tras el proceso de calibración. La cuarta incluye la cantidad de individuos de esa muestra cuyo nivel de conocimiento ha sido diagnosticado con un error de ± 1 . En la quinta columna se muestra el ECM total medido tras la calibración de las 200 CCO. Por último, en la sexta se representa el porcentaje de alumnos de la muestra utilizada a posteriori, cuyo nivel de conocimiento fue estimado adecuadamente.

En las tablas de la 7.11 a la 7.16 se presentan los datos que evidencian que el algoritmo de calibración es más preciso cuanto mayor es el tamaño de la muestra de examinandos que se utiliza en la calibración.

La figura 7.11 muestra que, para todos los casos, los mejores resultados se obtienen cuando el parámetro de suavizado es 0,8 o un valor cercano. Aunque no se ha representado gráficamente, a partir de las tablas se puede ver también que los mejores resultados, en cuanto a la inferencia del nivel de conocimiento de los individuos de la muestra empleada en la calibración, se obtienen también para el mismo valor del parámetro de suavizado.

7.7.3. Discusión

En este estudio se ha analizado cómo evolucionan los resultados del algoritmo de calibración propuesto, en función del tamaño de la muestra de alumnos de la muestra empleada para llevar a cabo la calibración, y del parámetro de suavizado. El objetivo era inferir 200 CCO de un total de 50 ítems de opción múltiple y de respuesta múltiple con opciones independientes.

Se ha mostrado como, incluso con sólo una muestra inicial de 20 alumnos, la evaluación que realizan los TAI que se administran a posteriori tiene una precisión aceptable (en torno al 90 % de individuos correctamente diagnosticados). De hecho, en estos tests a posteriori, cuando se infería erróneamente el nivel de conocimiento del examinando, el valor obtenido sólo difería en una unidad con respecto su nivel real.

A partir de esta información se puede concluir que a partir de este algoritmo de calibración se obtienen unos resultados satisfactorios, incluso con una muestra de alumnos pequeña. Esto permite disponer de un TAI aún cuando la información inicial de la que se dispone es de tamaño reducido.

7.7.4. Experimento 2: Estudio de la influencia de las funciones de suavizado en los resultados de la calibración

El objetivo de este experimento es estudiar qué función de suavizado, de las tres propuestas por Ramsay (véase sección 2.9.2), es la más adecuada para el algoritmo de calibración utilizado en el modelo de respuesta propuesto en esta tesis.

núm. alumnos	h	alumnos bien clasif.	alumnos mal clasif. a uno	ECM total	alumnos bien clasif. a posteriori
20	0,05	8	12	45,4514	100
20	0,1	8	11	40,2505	98
20	0,15	8	12	37,7685	100
20	0,2	8	12	36,4274	100
20	0,25	8	12	35,0987	100
20	0,3	8	12	33,5378	100
20	0,35	9	11	31,7083	100
20	0,4	12	8	29,7036	97
20	0,45	12	8	27,6584	100
20	0,5	13	7	25,7006	100
20	0,55	12	8	23,9389	100
20	0,6	14	6	22,4520	97
20	0,65	14	6	21,2909	98
20	0,7	13	7	20,4590	98
20	0,75	14	6	19,9167	97
20	0,8	14	6	19,6229	95
20	0,85	14	6	19,5461	95
20	0,9	13	7	19,6589	92
20	0,95	12	8	19,9419	97
20	1	12	8	20,3494	91
20	1,05	12	8	20,8435	90
20	1,1	12	8	21,4003	93
20	1,15	11	9	22,0003	83
20	1,2	11	9	22,6270	93
20	1,25	11	9	23,2670	89
20	1,3	11	8	23,9094	92
20	1,35	10	9	24,5457	93
20	1,4	9	9	25,1694	92
20	1,45	8	9	25,7759	92

Tabla 7.11: Resultados del proceso de calibración a partir de muestras de alumnos de diversos tamaños, utilizando distintos valores para el parámetro de suavizado (I parte).

núm. alumnos	h	alumnos bien clasif.	alumnos mal clasif. a uno	ECM total	alumnos bien clasif. a posteriori
50	0,05	15	30	43,5079	99
50	0,1	18	27	38,4833	100
50	0,15	20	27	36,1721	96
50	0,2	20	28	34,6718	98
50	0,25	20	28	32,9064	98
50	0,3	21	28	30,6734	99
50	0,35	33	17	28,2755	98
50	0,4	34	16	25,9770	98
50	0,45	36	14	23,8707	96
50	0,5	36	14	21,9670	100
50	0,55	35	15	20,2718	99
50	0,6	37	13	18,7958	100
50	0,65	37	13	17,5552	95
50	0,7	40	10	16,5616	98
50	0,75	41	9	15,8296	95
50	0,8	42	8	15,3663	98
50	0,85	43	7	15,1722	99
50	0,9	43	7	15,2172	100
50	0,95	41	9	15,4638	97
50	1	38	12	15,8774	98
50	1,05	38	12	16,4240	94
50	1,1	37	13	17,0691	92
50	1,15	36	14	17,7785	99
50	1,2	35	15	18,5250	95
50	1,25	34	16	19,2882	93
50	1,45	31	19	22,2628	88

Tabla 7.12: Resultados del proceso de calibración a partir de muestras de alumnos de diversos tamaños, utilizando distintos valores para el parámetro de suavizado (II parte).

núm. alumnos	h	alumnos bien clasif.	alumnos mal clasif. a uno	ECM total	alumnos bien clasif. a posteriori
100	0,05	34	66	39,9554	93
100	0,1	32	67	36,3390	97
100	0,15	32	67	34,3319	97
100	0,2	36	64	31,9783	100
100	0,25	39	61	28,8841	98
100	0,3	44	56	26,0622	97
100	0,35	58	42	23,7286	97
100	0,4	60	40	21,7213	95
100	0,45	53	47	19,9158	100
100	0,5	54	46	18,2786	96
100	0,55	57	43	16,8281	96
100	0,6	68	32	15,5957	99
100	0,65	93	7	14,6083	100
100	0,7	93	7	13,8899	100
100	0,75	91	9	13,4374	99
100	0,8	89	11	13,2291	98
100	0,85	88	12	13,2441	99
100	0,9	87	13	13,4593	97
100	0,95	87	13	13,8455	95
100	1	84	16	14,3682	99
100	1,05	77	23	14,9920	96
100	1,1	74	26	15,6866	93
100	1,15	70	30	16,4271	97
100	1,2	63	37	17,1940	94
100	1,25	65	35	17,9719	93
100	1,3	65	35	18,7487	94
100	1,35	67	33	19,5155	95
100	1,4	66	34	20,2655	94
100	1,45	63	37	20,9941	97

Tabla 7.13: Resultados del proceso de calibración a partir de muestras de alumnos de diversos tamaños, utilizando distintos valores para el parámetro de suavizado (III parte).

núm. alumnos	h	alumnos bien clasif.	alumnos mal clasif. a uno	ECM total	alumnos bien clasif. a posteriori
300	0,05	91	204	33,1606	100
300	0,1	110	188	31,9280	98
300	0,15	128	172	30,3587	99
300	0,2	137	163	27,8068	99
300	0,25	151	149	25,3291	97
300	0,3	163	137	23,0877	100
300	0,35	151	149	21,0515	97
300	0,4	152	148	19,2681	99
300	0,45	131	169	17,7495	98
300	0,5	73	227	16,4535	98
300	0,55	83	217	15,3406	97
300	0,6	94	206	14,4046	98
300	0,65	121	179	13,6649	98
300	0,7	210	90	13,1300	99
300	0,75	217	83	12,7938	97
300	0,8	236	64	12,6525	95
300	0,85	213	87	12,6982	96
300	0,9	202	98	12,9121	95
300	0,95	208	92	13,2713	98
500	0,05	236	258	33,5166	100
500	0,1	247	250	32,3047	100
500	0,15	262	236	30,0644	100
500	0,2	265	234	26,9071	100
500	0,25	282	217	23,8995	100
500	0,3	307	193	21,4316	99
500	0,35	308	192	19,4322	99
500	0,4	319	181	17,7244	97
500	0,45	101	399	16,2134	98
500	0,5	113	387	14,8826	98
500	0,55	276	224	13,7502	99
500	0,6	472	28	12,8463	98
500	0,65	462	38	12,1873	98
500	0,7	452	48	11,7729	99
500	0,75	430	70	11,5926	96
500	0,8	424	76	11,6300	99
500	0,85	396	104	11,8638	98
500	0,9	381	119	12,2682	98
500	0,95	359	141	12,8150	95

Tabla 7.14: Resultados del proceso de calibración a partir de muestras de alumnos de diversos tamaños, utilizando distintos valores para el parámetro de suavizado (IV parte).

núm. alumnos	h	alumnos bien clasif.	alumnos mal clasif. a uno	ECM total	alumnos bien clasif. a posteriori
1000	0,05	451	547	32,2921	99
1000	0,1	441	558	30,1779	99
1000	0,15	405	595	25,4280	99
1000	0,2	453	547	23,0982	100
1000	0,25	497	503	21,6714	100
1000	0,3	511	489	20,5248	100
1000	0,35	525	475	19,4457	95
1000	0,4	481	519	18,3595	94
1000	0,45	262	738	17,2862	94
1000	0,5	281	719	16,2771	94
1000	0,55	302	698	15,3761	95
1000	0,6	336	664	14,6116	97
1000	0,65	394	606	14,0022	97
1000	0,7	420	580	13,5596	96
1000	0,75	589	411	12,2857	99
1000	0,8	884	116	11,3727	99
1000	0,85	628	372	11,8092	97
1000	0,9	864	136	12,9821	98
1000	0,95	822	178	13,6768	97
5000	0,05	2448	2551	24,6096	99
5000	0,1	2082	2918	23,8583	100
5000	0,15	2122	2878	22,9952	100
5000	0,2	2317	2683	22,2511	97
5000	0,25	2720	2280	21,5304	95
5000	0,3	2846	2154	20,6561	96
5000	0,35	2877	2123	19,5481	97
5000	0,4	1402	3598	18,2896	99
5000	0,45	1325	3675	17,0472	96
5000	0,5	1498	3502	15,9309	100
5000	0,55	1653	3347	14,9772	96
5000	0,6	1837	3163	14,2001	98
5000	0,65	1854	3146	13,6142	100
5000	0,7	1893	3107	13,2271	95
5000	0,75	4758	242	12,0304	98
5000	0,8	4895	105	10,0061	100
5000	0,85	4521	479	11,1359	95
5000	0,9	4367	633	12,4026	98
5000	0,95	3750	1250	13,7896	99

Tabla 7.15: Resultados del proceso de calibración a partir de muestras de alumnos de diversos tamaños, utilizando distintos valores para el parámetro de suavizado (V parte).

núm. alumnos	h	alumnos bien clasif.	alumnos mal clasif. a uno	ECM total	alumnos bien clasif. a posteriori
10000	0,05	3533	6467	26,5161	100
10000	0,1	4310	5690	25,3110	100
10000	0,15	4300	5700	23,0486	100
10000	0,2	4917	5083	29,8900	100
10000	0,25	5861	4139	18,5775	97
10000	0,3	5822	4178	16,5034	99
10000	0,35	5763	4237	14,8166	99
10000	0,4	2612	7388	13,4287	89
10000	0,45	2692	7308	13,2310	94
10000	0,5	3009	6991	13,1718	96
10000	0,55	3372	6628	11,2476	96
10000	0,6	3500	6500	11,1843	99
10000	0,65	3673	6327	10,9108	97
10000	0,7	3740	6260	10,5368	98
10000	0,75	9712	288	10,3509	100
10000	0,8	9646	354	9,3343	100
10000	0,85	8571	1429	12,4702	99
10000	0,9	4935	5065	13,7436	97
10000	0,95	6165	3835	14,1385	96

Tabla 7.16: Resultados del proceso de calibración a partir de muestras de alumnos de diversos tamaños, utilizando distintos valores para el parámetro de suavizado (VI parte).

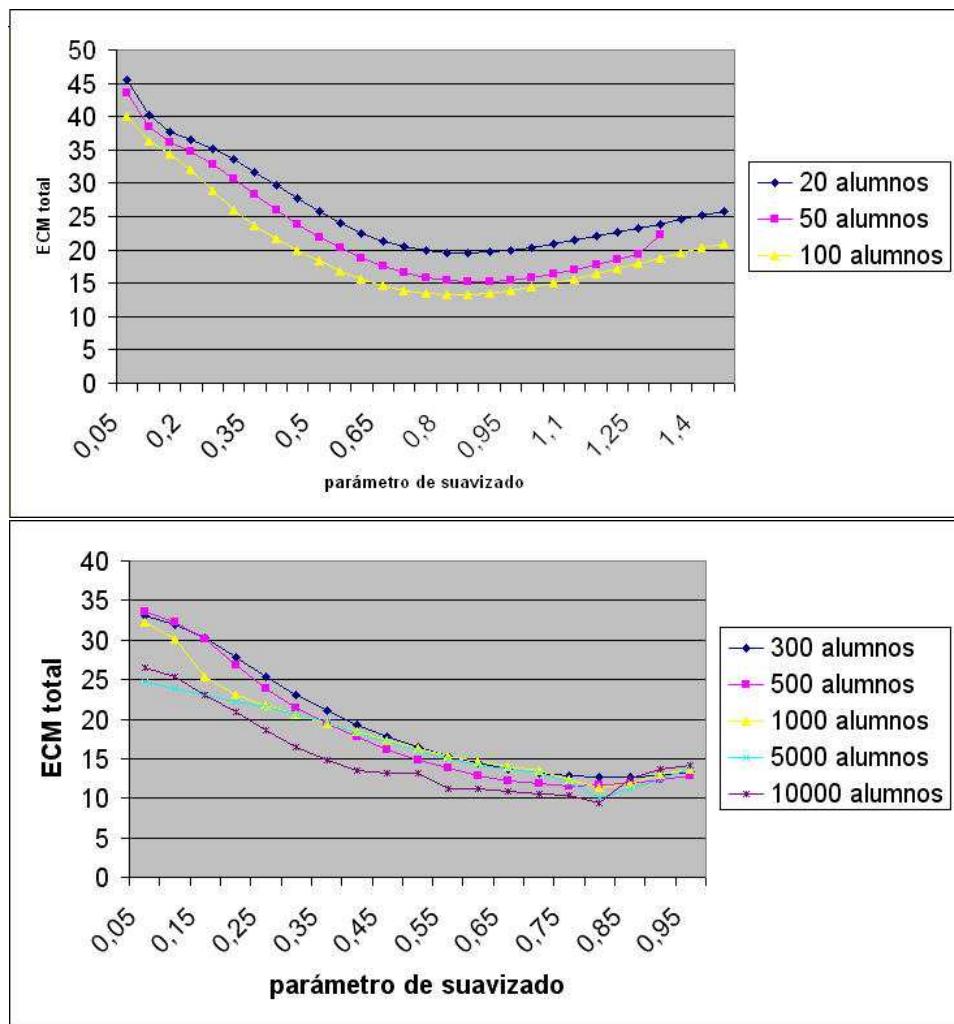


Figura 7.11: Resultados de la calibración llevada a cabo con muestras de alumnos de diversos tamaños, a partir de valores diferentes del parámetro de suavizado.

Para este estudio se ha utilizado un banco formado por un total de 50 ítems. Asimismo, se ha sometido a una población de 300 alumnos simulados a un test con todos esos ítems. Posteriormente se han llevado a cabo diversas calibraciones variando, por un lado el valor del parámetro de suavizado, y por otro, modificando la función de suavizado empleada. Así, por cada valor del parámetro, se han llevado tres calibraciones con las tres funciones núcleo propuestas por Ramsay; esto es, función epanechnikov, gaussiana y uniforme. Las inferencias de las CCO se han realizado tomando como rango, y por tanto, como escala para medir el nivel de conocimiento, un valor igual a seis. Los niveles reales de los alumnos fueron generados siguiendo una distribución normal. Asimismo, las CCO de los ítems fueron creadas siguiendo funciones 3PL. Los valores del factor de discriminación, dificultad y adivinanza fueron generados siguiendo distribuciones normales centradas en los valor 1, 2, 2 y 0,15, respectivamente.

Tras cada proceso de calibración, se ha sometido a una población de 100 nuevos examinandos virtuales a un TAI con el banco de ítems recién calibrado. Los niveles de conocimiento reales de estos alumnos fueron generados también siguiendo una distribución normal. Los tests administrados tenían las siguientes características: utilizaban MAP como criterio de evaluación; el método de selección de ítems era el basado en la entropía; y el criterio de finalización el basado en la dispersión con un umbral igual a 0,001.

7.7.5. Resultados

Las tablas 7.17 y 7.18 muestran la información resultante de las diferentes calibraciones realizadas. Obsérvese que el valor inferior del parámetro de suavizado (primera columna) es 0,25. Esto es debido a que las calibraciones realizadas empleando las funciones epanechnikov y uniforme, para valores menores que 0,25, dan como resultado inconsistencias.

Cada fila de la tabla representa un proceso diferente de calibración de los ítems. La primera columna indica el parámetro de suavizado utilizado; la segunda el tipo de función núcleo; la tercera el número de alumnos de la muestra empleada en la calibración, para los cuales, tras finalizar ésta, su nivel de conocimiento ha sido inferido correctamente. En la cuarta columna se representa el número de individuos cuyo nivel inferido difiere tan sólo en una unidad (bien sea por encima, o por debajo) con respecto a su valor real. La quinta contiene el ECM de las CCO calibradas con respecto a las originales. Este valor es un indicador global para todas las CCO. Por último, la sexta columna indica el número de alumnos (de la muestra utilizada para la validación a posteriori), cuyo nivel ha sido estimado correctamente utilizando un TAI cuyo banco estaba formado tan sólo por los ítems recién calibrados.

En la figura 7.12 se han incluido dos gráficos en los que se representan los resultados de la calibración en términos de la función de suavizado utilizada. En el eje de abscisas se representan los valores del parámetro de suavizado. El de ordenadas contiene, en el gráfico superior, el número de alumnos; y en el inferior el ECM total. Cada función representada en el gráfico superior, muestra como varía el número de alumnos (de la muestra utilizada para la calibración) clasificados correctamente dependiendo del valor del parámetro de suavizado, para una función núcleo concreta. Como se puede apreciar, la función núcleo con la que se obtiene un número mayor de alumnos mejor clasificado es la gaussiana (para valores cercanos a 0,75 del parámetro de suavizado), seguida muy de cerca por la epanechnikov.

En la gráfica inferior se está representando como varía el ECM total en función del parámetro de suavizado. En este caso, también se puede apreciar que la función núcleo con

<i>h</i>	f. núcleo	alumnos bien clasif.	alumnos mal clasif. a uno	ECM total	alumnos bien clasif. a posteriori
0,25	epach.	131	163	25,56	100
0,25	gauss.	230	70	14,27	99
0,25	unif.	138	156	24,98	100
0,3	epach.	127	165	25,84	100
0,3	gauss.	250	50	12,17	100
0,3	unif.	126	166	26,02	100
0,35	epach.	132	163	25,66	100
0,35	gauss.	259	41	10,35	99
0,35	unif.	132	163	25,66	100
0,4	epach.	136	162	21,17	99
0,4	gauss.	262	38	9,29	100
0,4	unif.	137	161	19,30	100
0,45	epach.	143	155	19,03	98
0,45	gauss.	270	30	8,75	100
0,45	unif.	145	153	18,92	100
0,5	epach.	155	143	18,31	99
0,5	gauss.	273	27	8,62	97
0,5	unif.	155	143	18,31	99
0,55	epach.	156	142	18,26	100
0,55	gauss.	283	17	7,85	98
0,55	unif.	163	135	16,65	100
0,6	epach.	163	135	17,78	100
0,6	gauss.	284	16	7,14	98
0,6	unif.	175	123	15,91	100
0,65	epach.	185	115	15,30	100
0,65	gauss.	283	28	6,89	97
0,65	unif.	190	110	15,04	100
0,7	epach.	187	113	15,22	99
0,7	gauss.	288	12	6,759	96
0,7	unif.	213	87	12,22	100
0,75	epach.	193	107	14,34	100
0,75	gauss.	289	11	6,62	97
0,75	unif.	218	82	12,02	100

Tabla 7.17: Resultados de utilizar diferentes funciones núcleo para la calibración de un banco de 50 ítems (I parte).

h	f. núcleo	alumnos bien clasif.	alumnos mal clasif. a uno	ECM total	alumnos bien clasif. a posteriori
0,8	epach.	200	100	13,54	99
0,8	gauss.	275	25	7,73	95
0,8	unif.	231	69	10,86	98
0,85	epach.	229	71	11,37	100
0,85	gauss.	261	39	8,62	98
0,85	unif.	235	65	10,31	100
0,9	epach.	237	63	10,49	100
0,9	gauss.	226	74	9,24	97
0,9	unif.	245	55	9,97	100
0,95	epach.	241	59	10,39	100
0,95	gauss.	212	88	10,04	98
0,95	unif.	255	45	9,527	99
1	epach.	249	51	9,93	100
1	gauss.	199	101	11,67	99
1	unif.	225	75	11,80	96
1,05	epach.	286	14	6,82	100
1,05	gauss.	195	105	12,09	94
1,05	unif.	225	75	11,80	96
1,1	epach.	288	12	6,68	99
1,1	gauss.	192	108	13,54	96
1,1	unif.	225	75	11,80	93
1,15	epach.	284	16	7,05	98
1,15	gauss.	188	112	14,00	97
1,15	unif.	225	75	11,80	95
1,2	epach.	278	22	7,56	98
1,2	gauss.	186	114	14,47	93
1,2	unif.	225	75	11,80	97
1,25	epach.	272	28	8,00	97
1,25	gauss.	184	116	14,97	98
1,25	unif.	225	75	11,80	97
1,3	epach.	267	33	8,36	99
1,3	gauss.	181	119	15,49	94
1,3	unif.	225	75	11,80	97
1,35	epach.	262	38	8,71	96
1,35	gauss.	176	124	15,94	94
1,35	unif.	225	75	11,80	97
1,4	epach.	262	38	8,87	96
1,4	gauss.	173	127	16,47	96
1,4	unif.	225	75	11,80	94
1,45	epach.	259	41	9,09	98
1,45	gauss.	170	130	17,00	94
1,45	unif.	225	75	11,80	95

Tabla 7.18: Resultados de utilizar diferentes funciones núcleo para la calibración de un banco de 50 ítems (II parte).

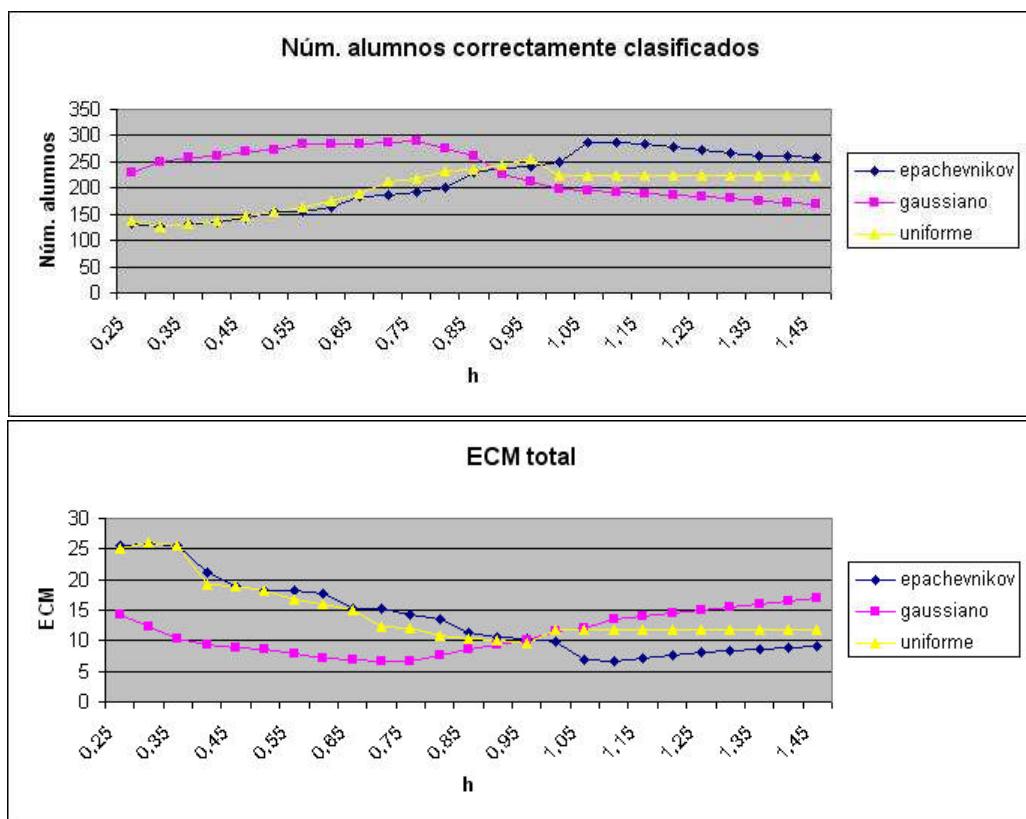


Figura 7.12: Resultados de la calibración cuando se aplican diferentes funciones núcleo:
Arriba: número de alumnos clasificados correctamente. Abajo: error cuadrático medio total.

la que se obtiene mejores resultados es la gaussiana (en el entorno del valor del parámetro de suavizado 0,75), seguida también esta vez muy de cerca por la epanechnikov.

7.7.6. Discusión

Además de las simulaciones anteriormente descritas, se han realizado otros experimentos en condiciones diferentes. En todos los casos, se ha puesto de manifiesto que la función núcleo con la que se obtienen mejores resultados, en el algoritmo de calibración propuesto en esta tesis, es la función gaussiana. Aunque es cierto, que en casos como el expuesto, los resultados obtenidos con la función epanechnikov (aunque menores) se acercaban bastante.

7.7.7. Experimento 3: Estudio de los heurísticos utilizados para la calibración

Este experimento trata de demostrar que, para la calibración de bancos en los que se incluyen ítems de respuesta múltiple con opciones independientes, es más adecuado utilizar el criterio heurístico de evaluación por puntos para la primera fase del algoritmo de calibración, esto es, para la fase de cálculo de las evaluaciones.

Obviamente, no existe ninguna diferencia si el banco está exclusivamente formado por ítems de opción múltiple, ya que en esta ocasión, los criterios heurísticos de evaluación porcentual y por puntos son equivalentes. Por este motivo, se han llevado a cabo diversas simulaciones con un banco formado exclusivamente por ítems de respuesta múltiple con opciones independientes.

El banco utilizado se compone de 50 ítems con cuatro opciones de respuesta cada uno. Para llevar a cabo la simulación se han utilizado 300 alumnos simulados, cuyos niveles de conocimiento reales han sido generados siguiendo una distribución normal. El rango de todas las curvas características, y por tanto la escala de niveles de conocimiento está compuesta por 6 niveles (de 0 a 5). Las CCO de los ítems han sido generados de forma que, para que uno de esos ítems sea evaluado como completamente correcto, el alumno debe seleccionar dos de las opciones (las dos primeras), y dejar sin seleccionar las dos restantes. Todas las curvas han sido creadas siguiendo el clásico modelo 3PL dicotómico. Los valores del factor de discriminación, la dificultad y la adivinanza fueron generados siguiendo distribuciones normales centradas en 1,2, 2 y 0,15, respectivamente.

7.7.8. Resultados

Una vez simulada la realización del test por parte de los 300 estudiantes virtuales, a partir de la información resultante, se procedió a la calibración de los ítems. Ésta ha sido realizada utilizando la función núcleo gaussiana. Se han llevado a cabo diversas calibraciones en las que los parámetros que han variado han sido el ancho de banda, h (parámetro de suavizado); y el criterio heurístico aplicado a la primera fase de la calibración.

Para verificar la bondad de la calibración, una vez calibradas las CCO se sometía a una nueva población de 100 individuos a un TAI en el que el banco estaba formado por los 50 ítems calibrados. En éste el criterio de evaluación era el MAP; el mecanismo de selección de ítems el basado en la entropía; y por último, el método de finalización, el basado en la dispersión con un umbral de 0,001.

<i>h</i>	heurístico	alumnos bien clasif.	alumnos mal clasif. a uno	ECM total	alumnos bien clasif. a posteriori
0,45	porcent.	30	246	83,74	98
0,45	puntos	193	88	65,23	100
0,5	porcent.	30	264	80,09	100
0,5	puntos	205	81	60,52	100
0,55	porcent.	30	270	78,58	99
0,55	puntos	214	81	56,55	100
0,6	porcent.	30	270	78,57	99
0,6	puntos	273	27	40,37	100
0,65	porcent.	31	269	75,97	97
0,65	puntos	272	28	38,17	100
0,7	porcent.	31	269	75,32	100
0,7	puntos	272	28	29,57	100
0,75	porcent.	32	268	72,56	100
0,75	puntos	280	20	17,76	100
0,8	porcent.	32	268	69,22	100
0,8	puntos	281	19	16,19	100
0,85	porcent.	44	256	62,32	100
0,85	puntos	249	51	32,46	95
0,9	porcent.	82	218	54,66	98
0,9	puntos	229	71	37,57	98

Tabla 7.19: Resultados de aplicar distintos criterios heurísticos para la calibración de un banco de 50 ítems.

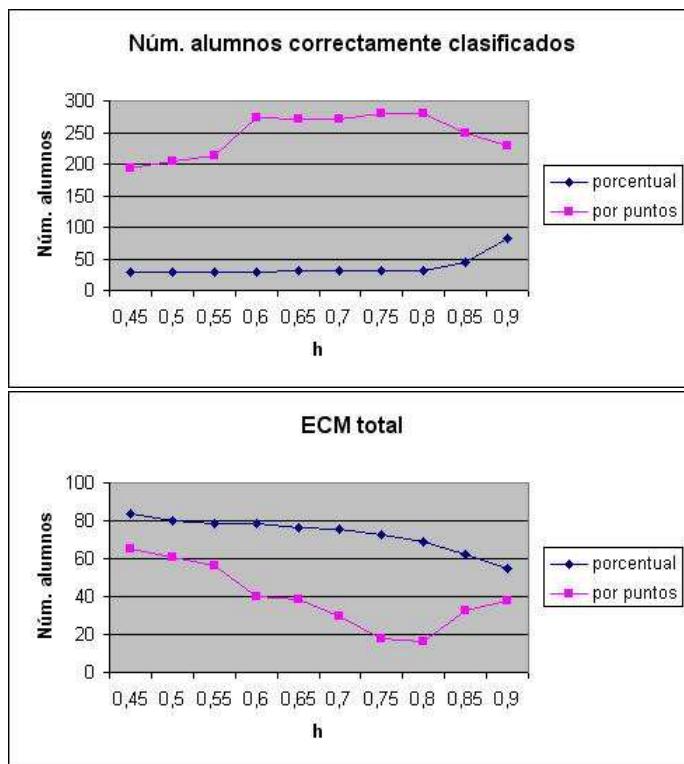


Figura 7.13: Resultados de la calibración cuando se aplican los dos criterios heurísticos de evaluación: Arriba: número de alumnos clasificados correctamente. Abajo: error cuadrático medio total tras la calibración.

La tabla 7.19 muestra los resultados por filas de las distintas calibraciones. La primera columna contiene el valor del parámetro de suavizado que fue aplicado en esa calibración. La segunda el criterio heurístico utilizado en la primera fase del procedimiento de calibración. Sus valores pueden ser: (a) *porcent.*, que indica que se aplicó el criterio de evaluación porcentual; (b) *puntos*, cuando el método empleado fue el por puntos. En este último caso, a cada opción de respuesta se le asignaban 0,25 puntos; de forma que, si todas las opciones estaban seleccionadas adecuadamente el ítem recibía un punto.

La tercera columna de la tabla 7.19 representa el número de alumnos de la muestra utilizada para realizar la calibración que, tras aplicar la fase quinta del algoritmo, su nivel de conocimiento fue inferido correctamente. En la cuarta se han incluido también la cantidad de individuos de la muestra cuyo nivel fue determinado erróneamente, pero con un error de una unidad. La quinta columna recoge el ECM total de todas las CCO (las 200 curvas) tras la calibración. Por último, la sexta contiene el número de alumnos bien clasificados en la muestra utilizada a posteriori para validar los resultados de la calibración.

Como se puede apreciar en la tabla, las diferencias entre ambos criterios heurísticos son sustanciales cuando se compara el número de alumnos evaluados correctamente; y cuando se tiene en cuenta el ECM total obtenido. Ciertamente, no existen grandes diferencias en lo referente al diagnóstico a posteriori mediante TAI con las curvas calibradas.

Para poder apreciar mejor las diferencias en la aplicación de ambos criterios heurísticos, se han incluido dos gráficas (figura 7.13). En la superior se ha representado la evolución del número de alumnos de la muestra correctamente clasificados tras la calibración, en función del valor del parámetro de suavizado empleado. Por otra parte, la gráfica inferior recoge la evolución, también según el parámetro de suavizado aplicado, del ECM total de las CCO, tras la calibración.

7.7.9. Discusión

Los resultados del experimento anterior han puesto de manifiesto que el uso del criterio heurístico de evaluación por puntos parece ser el más adecuado cuando se calibran ítems de crédito parcial, como es el caso de los de respuesta múltiple con opciones independientes. Por el contrario, el criterio porcentual no es capaz de distinguir, en ítems en los que se permiten respuestas parcialmente correctas, entre este caso y aquél en el que el examinando responde de forma completamente incorrecta.

La aplicación de la primera fase del algoritmo de calibración es necesaria para poder ordenar a los alumnos según sus resultados en el test, para así poder acometer la fase de suavizado. Por este motivo, cuando en el test intervienen ítems que permiten crédito parcial, si se aplica el criterio de evaluación porcentual, la ordenación resultante puede no ser del todo correcta, tal y como se ha podido ver en este estudio.

7.7.10. Experimento 4: Comparación del algoritmo de calibración con la propuesta original de Ramsay

El objetivo de este experimento es comparar el algoritmo de calibración basado en el suavizado núcleo, en su versión original propuesta por Ramsay, con la modificación que se ha llevado a cabo en esta tesis. Con este fin, se han realizado diversas simulaciones comparativas con una muestra poblacional de 300 alumnos simulados. Sus niveles de conocimiento reales fueron generados siguiendo una distribución normal.

Estos alumnos realizaron un test heurístico compuesto por 50 ítems cuyas CCO se deseaban calibrar. Siguiendo el procedimiento habitual, las curvas reales de esos ítems fueron generadas siguiendo el modelo 3PL, donde los valores del factor de discriminación, la dificultad y la adivinanza se generarán siguiendo distribuciones normal centradas en 1, 2, 2 y 0,15, respectivamente. El número de niveles de evaluación, y por tanto, el rango de las CCO está formado por seis valores.

7.7.11. Resultados

A partir de las sesiones realizadas por los 100 alumnos, se llevaron a cabo diversas calibraciones en las que se modificó el algoritmo utilizado. En unos casos, ésta se realizó siguiendo la propuesta original de Ramsay, y en otros, la modificación realizada para el modelo de respuesta presentado en esta tesis. A la par también se fue variando el valor del parámetro de suavizado empleado.

Nótese que, por tratarse de un banco mixto formado en partes iguales por ítems de opción múltiple con tres opciones de respuesta, y de respuesta múltiple con cuatro opciones independientes, como heurístico aplicado a la primera fase de la calibración se ha utilizado el criterio por puntos. Recuérdese que éste permite mejores resultados que el porcentual. Ha

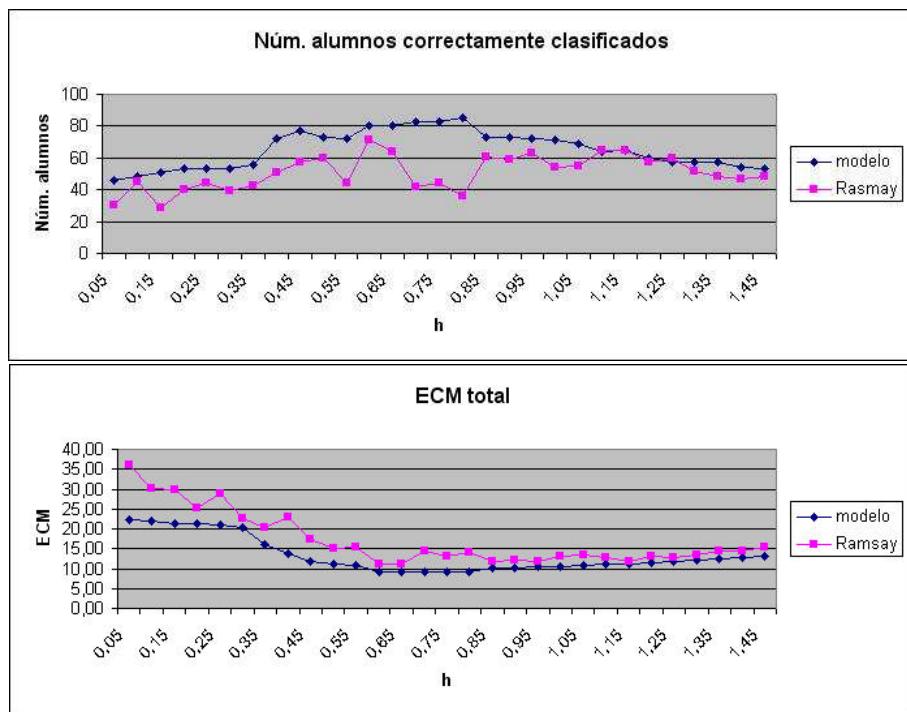


Figura 7.14: Resultados de la calibración cuando se aplican las dos versiones del algoritmo. Arriba: número de alumnos clasificados correctamente. Abajo: error cuadrático medio total tras la calibración.

sido empleado, por tanto, no sólo al algoritmo de calibración del modelo, sino también a la propuesta original de Ramsay. En ambos casos, se ha hecho uso de la función de suavizado gaussiana.

Las tablas 7.20 y 7.21 son análogas a las de los experimentos anteriores. La primera columna contiene el parámetro de suavizado empleado; la segunda, el algoritmo de calibración utilizado en cada caso (*Ramsay* indica la propuesta original de Ramsay, *modelo* la versión del algoritmo desarrollada en esta tesis); la tercera el número de alumnos de la muestra inicial cuyo nivel de conocimiento fue diagnosticado correctamente; la cuarta, la cantidad de individuos cuyo nivel fue diagnosticado con un error igual a una unidad; la quinta columna el ECM total de las CCO de todos los ítems; y por último, la sexta el porcentaje de examinandos cuyo nivel fue diagnosticado correctamente en los TAI de validación de las curvas calibradas.

En la figura 7.14 se muestran dos gráficas comparativas de los resultados después de aplicar las dos versiones del algoritmo de calibración basado en el suavizado núcleo. En la gráfica superior se ha representado como evoluciona el número de alumnos (de la muestra original) diagnosticados correctamente tras la calibración, en función del valor del parámetro de suavizado utilizado en cada caso. Como se puede apreciar, los mejores resultados se obtiene cuando se aplica la versión del algoritmo desarrollada para el modelo de respuesta de la tesis, para el valor del parámetro de suavizado 0,8. En este caso, tan sólo seis de los alumnos de la muestra fueron diagnosticados erróneamente, y además (como muestran las

h	método calibrac.	alumnos bien clasif.	alumnos mal clasif. a uno	ECM total	alumnos bien clasif. a posteriori
0,05	modelo	46	52	22,37	97
0,05	Ramsay	30	26	36,08	71
0,1	modelo	48	51	22,08	97
0,1	Ramsay	45	32	30,03	84
0,15	modelo	51	49	21,46	98
0,15	Ramsay	29	48	29,99	85
0,2	modelo	53	47	21,22	97
0,2	Ramsay	40	49	25,38	97
0,25	modelo	53	47	21,13	98
0,25	Ramsay	44	44	28,72	79
0,3	modelo	53	47	20,30	99
0,3	Ramsay	39	54	22,56	80
0,35	modelo	56	44	16,04	97
0,35	Ramsay	43	57	20,37	88
0,4	modelo	72	28	13,82	100
0,4	Ramsay	51	43	22,84	97
0,45	modelo	77	23	11,69	99
0,45	Ramsay	57	37	17,26	98
0,5	modelo	73	27	11,13	98
0,5	Ramsay	60	39	15,16	96
0,55	modelo	72	28	10,72	98
0,55	Ramsay	44	54	15,31	91
0,6	modelo	80	20	9,27	98
0,6	Ramsay	71	29	11,07	93
0,65	modelo	80	20	9,31	92
0,65	Ramsay	64	36	11,22	87
0,7	modelo	83	17	9,05	96
0,7	Ramsay	42	58	14,55	96
0,75	modelo	83	17	9,16	91
0,75	Ramsay	44	56	13,19	93
0,8	modelo	85	15	9,04	99
0,8	Ramsay	36	64	14,25	91
0,85	modelo	73	27	10,13	87
0,85	Ramsay	61	39	11,81	81
0,9	modelo	73	27	10,20	92
0,9	Ramsay	59	41	12,18	86
0,95	modelo	72	28	10,37	89
0,95	Ramsay	63	37	11,87	85

Tabla 7.20: Resultados de la calibración aplicando el algoritmo de suavizado de Ramsay y la versión utilizada en el modelo de respuesta (I parte).

h	método calibrac.	alumnos bien clasif.	alumnos mal clasif. a uno	ECM total	alumnos bien clasif. a posteriori
1	modelo	71	29	10,57	89
1	Ramsay	54	41	13,18	84
1,05	modelo	69	31	10,77	87
1,05	Ramsay	55	41	13,43	77
1,1	modelo	64	36	11,14	78
1,1	Ramsay	65	33	12,84	80
1,15	modelo	65	35	11,28	88
1,15	Ramsay	65	35	11,83	74
1,2	modelo	60	40	11,62	91
1,2	Ramsay	57	40	13,06	72
1,25	modelo	57	43	11,92	80
1,25	Ramsay	60	37	12,66	74
1,3	modelo	57	43	12,18	80
1,3	Ramsay	52	43	13,52	71
1,35	modelo	57	42	12,43	80
1,35	Ramsay	48	47	14,28	70
1,4	modelo	54	44	12,67	73
1,4	Ramsay	47	48	14,48	69
1,45	modelo	53	43	12,97	77
1,45	Ramsay	48	47	15,26	59

Tabla 7.21: Resultados de la calibración aplicando el algoritmo de suavizado de Ramsay y la versión utilizada en el modelo de respuesta (II parte).

tablas 7.20 y 7.21) el error cometido era de un solo nivel de conocimiento. Según Ramsay, con su algoritmo se deberían haber obtenido los mejores resultados para el valor del parámetro de suavizado igual a 0,35, lo cual no se corresponde con los resultados.

En la gráfica inferior de la figura 7.14, se ha representado la evolución del error cuadrático medio total de todas las CCO, en función del valor del parámetro de suavizado. Como se puede apreciar, en este caso, los mejores resultados se obtienen también con la versión del algoritmo desarrollada para el modelo de respuesta presentado. El mejor resultado se obtiene también para el valor del parámetro de suavizado 0,8.

7.7.12. Discusión

Como se ha puesto de manifiesto a través de este experimento, las modificaciones realizadas sobre el algoritmo de calibración de modelos de respuesta no paramétricos, basado en el suavizado núcleo, se ha traducido en una mejora de los resultados que se obtienen. También es necesario decir que la nueva versión de ese algoritmo presenta una desventaja con respecto a su predecesor: el refinamiento iterativo requiere de un número mayor de iteraciones que la propuesta original de Ramsay. Mientras que la propuesta original de Ramsay, tal y como él mencionaba (Ramsay, 2000), requiere de un número menor o igual a tres iteraciones; la nueva versión desarrollada puede requerir, en los casos más extremos, incluso once iteraciones. Ciertamente, esto no es problemático desde el punto de vista computacional, puesto que el tiempo de computo requerido para ese número de iteraciones es de poco más de un minuto (en un PC con un procesador Pentium IV a 2GHz.).

7.7.13. Conclusiones del estudio

En este estudio se han puesto de manifiesto las características del algoritmo de calibración propuesta en esta tesis. En el primero de los experimentos se ha estudiado cómo se comporta el algoritmo frente a muestras de alumnos de diferentes tamaños, y a su vez a diversos parámetros de suavizado. Como era de esperar, cuanto mayor es el tamaño de los resultados utilizados para la calibración, se obtiene estimaciones más precisas de las CCO. Asimismo, se ha mostrado también que, incluso con un conjunto reducido de alumnos iniciales (20 alumnos), las CCO que se obtienen permiten la administración de TAI, cuyos diagnósticos tienen un porcentaje de error aceptable.

En el segundo experimento se han estudiado las tres funciones de suavizado propuestas por Ramsay, para determinar con cuál de ellas se obtienen los mejores resultados. La conclusión del experimento es que con la función gaussiana se obtienen los mejores resultados, seguida de cerca por la función epanechnikov.

El tercer experimento aborda cómo afectan los heurísticos empleados en la primera fase del algoritmo de calibración, a sus resultados finales. El resultado es que utilizando el criterio por puntos, se obtienen mejores estimaciones de las CCO de los ítems que se desea calibrar, cuando entre esos ítems se encuentran ítems que permiten créditos parciales, como por ejemplo, los ítems de respuesta múltiple con opciones independientes.

Por último, el cuarto experimento ha comparado el algoritmo propuesto con la versión inicial propuesta por Ramsay. Los resultados muestran que, al menos para el modelo de respuesta propuesto en esta tesis, con el algoritmo propuesto se obtienen mejores resultados.

7.8. Estudio de la viabilidad de SIETTE como medio de recolección de evidencias para la calibración de ítems

En esta sección se recoge un estudio cuyo objetivo era demostrar empíricamente que un sistema de tests a través de la Web (en este caso SIETTE) es válido como medio de recolección de datos para la calibración de ítems (Guzmán et al., 2002). Como se ha descrito en el capítulo anterior, SIETTE permite administrar no sólo TAI, sino también tests convencionales. Incluye para este tipo de tests dos criterios de selección de evaluación heurísticos: el criterio por puntos y el porcentual.

La administración de tests a través de papel y lápiz es un método cuya validez como medio de recolección de información para la calibración de ítems es comúnmente reconocida. Ciertamente, la administración a través de la Web presenta ciertos inconvenientes. La ubicuidad de la Web imposibilita tener control sobre las condiciones en las que los alumnos están realizando el test. Por este motivo, al igual que ocurre con los tests de lápiz y papel, si un test se administra a través de la Web sin la supervisión adecuada, en principio, no puede recopilar evidencias válidas para la calibración de los ítems.

El objetivo de este estudio era de tipo comparativo. La idea era llevar a cabo una comparación entre la evaluación realizada mediante tests administrados mediante papel y lápiz, y tests administrados con SIETTE. El objetivo era mostrar que la recolección de datos mediante ambos métodos es equivalente. Esto afianzaría la idea de que, mediante un sistema Web, es posible calibrar un conjunto de ítems de forma correcta.

En este estudio se compararon los resultados obtenidos en un test administrado a través de SIETTE con los resultados obtenidos de administrar ese mismo test a través de papel y lápiz. Ambos tests compartían el mismo número de ítems, el criterio de selección de ítems, el formato de los ítems, el criterio de finalización del test, y el tiempo que disponían los alumnos para completarlo. SIETTE fue utilizado como medio de administración del test a través de la Web, y por tanto, como medio de recolección de datos. El test, que evaluaba un único concepto, se componía de veinte ítems sobre gramática inglesa, todos ellos de opción múltiple con tres opciones de respuesta. El alumno disponía de un tiempo máximo de veinte minutos para completarlo. El criterio de evaluación utilizado fue el porcentaje de ítems respondidos correctamente. Los ítems fueron administrados siempre (para todos los alumnos) en el mismo orden. Previamente a la administración del test, a cada alumno se le mostraba una página en la que se explicaba el experimento del que iba a formar parte. Inmediatamente después, se pedía a los alumnos que llenaran un cuestionario personal. En este cuestionario se le hacían preguntas sobre su edad, sexo, nivel de estudios (con las opciones: BUP o ESO, COU o bachillerato, universitario de primer ciclo, universitario de segundo ciclo, licenciatura en filología inglesa, otra licenciatura o doctorado), así como una estimación personal sobre su nivel de inglés (medido mediante una escala Likert de 1 a 5: prácticamente nulo, bajo, medio, alto o prácticamente bilingüe), el origen de su formación en inglés, su nacionalidad y su lengua materna. Otra información recopilada de forma automática e interna por SIETTE eran la fecha, tiempo de conexión, sistema operativo y dirección IP del PC desde el que estaban realizando el test, herramienta software de navegación utilizada, y finalmente el tiempo de exposición de cada ítem. Tras llenar y enviar este cuestionario, se mostraban las instrucciones que el alumno tenía que tener en cuenta para la realización del test. A partir del momento en el que se mostraba el primer ítem al alumno, el tiempo disponible para la realización del test comenzaba a decrementarse. Una vez finalizado el test, al alumno

se le mostraba únicamente su puntuación final (porcentaje de respuestas correctas). Al final se incluía un formulario para enviar sugerencias y/o comentarios con respecto al test.

7.8.1. Resultados

Los alumnos que participaron en el experimento fueron reclutados utilizando diversos medios: emails, panfletos de propaganda y carteles colocados principalmente en centros universitarios. En el estudio participaron investigadores de tres universidades españolas: la Universidad Politécnica de Valencia, la Universidad Autónoma de Madrid y la Universidad de Málaga. Se recopilaron datos durante aproximadamente unos dos meses (entre el 1 de abril y el 15 de mayo de 2002). Un total de 2316 sesiones del test fueron recopiladas tal y como se muestra en la tabla 7.22. Tras un análisis exhaustivo de los datos, realizado por los psicómetras que participaron en el estudio, sólo 1139 sesiones fueron consideradas válidas. Los principales motivos para descartar los datos restantes fueron la existencia de sesiones completas en blanco (el estudiante no llegó a realizar el test); comentarios finales de los alumnos que podrían indicar falta de rigor en la realización del test; tests incompletos con demasiadas omisiones y tiempos de exposición excesivamente bajos; posibles repeticiones en la realización del test por parte de un mismo alumno (fueron detectadas por la IP del PC desde la que se realizó la conexión con SIETTE); o bien tests realizados por individuos cuya lengua materna no era el español. Obsérvese que, aunque en este caso el filtrado se hizo de forma manual, la mayoría de los criterios que se aplicaron podrían ser fácilmente automatizados.

	Núm. alumnos
Tests en blanco	474
Comentarios	28
Tests incompletos, omisiones o tiempo	322
Misma dirección IP	325
Lengua materna no español	28
Total datos descartados	1177
Total datos utilizados	1139
Total recopilado	2316

Tabla 7.22: Datos recopilados a través de SIETTE.

Los alumnos que realizaron el test procedían principalmente de España y de otros países de habla hispana, tal y como se muestra en la figura 7.16. En el test administrado mediante papel y lápiz, participaron un total de 435 personas, todas ellas de nacionalidad española. En cuanto al sexo de los alumnos, la muestra poblacional que realizó el test a través de la Web, sólo un 39 % eran hombres, y en el test administrado mediante lápiz y papel la cifra de individuos masculinos se reducía al 24 %.

En lo referente al nivel de inglés, ambas muestras poblacionales daban unas proporciones similares tal y como se muestra en la gráfica superior de la figura 7.17. En esta figura se describe la proporción de alumnos de cada muestra según su nivel de inglés estimado por ellos mismos. En cuanto a la edad y los estudios, las diferencias entre ambas muestras poblacionales son significativas como se puede apreciar en las gráficas central e inferior de la figura 7.17. Mientras que los individuos que realizaron el test mediante lápiz y papel eran

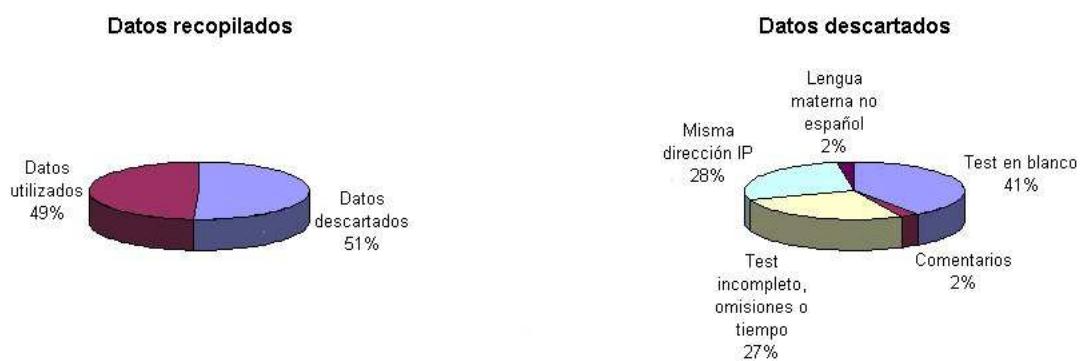


Figura 7.15: Porcentajes de datos recopilados y descartados.



Figura 7.16: Distribución de alumnos según su nacionalidad.

principalmente estudiantes de primer ciclo universitario, para el test administrado mediante SIETTE las proporciones son más similares.

Como muestra la gráfica inferior de la figura 7.17, las medias de edades de ambas muestras eran también similares. En el test de papel y lápiz la media era de 20,2 años, y en el test a través de la Web de 25,9. A pesar de este dato, el rango de edades de los individuos que realizar el test en SIETTE era más amplio (entre 12 y 65 años) que en el administrado por lápiz y papel (entre 15 y 48 años).

Los datos fueron comparados estadísticamente con los recopilados en el test administrado mediante papel y lápiz. Asimismo uno de los ítems fue descartado porque se descubrió que había sido redactado erróneamente. En cuanto a las propiedades psicométricas de los datos, aunque el factor de discriminación calibrado variaba entre ambas muestras en un máximo de 0,16 para el ítem 15 (figura 7.18), la dificultad de los ítems era bastante similar en ambos entornos. La diferencia máxima se daba en el ítem 7 y era de 0,1. Asimismo, si se analizan las estimación de los parámetros de las CCI de forma global, se puede concluir que los resultados obtenidos a través de ambos medios de recolección de evidencias eran bastante similares.

Otra forma de comparar los resultados de ambas calibraciones de forma general es ana-

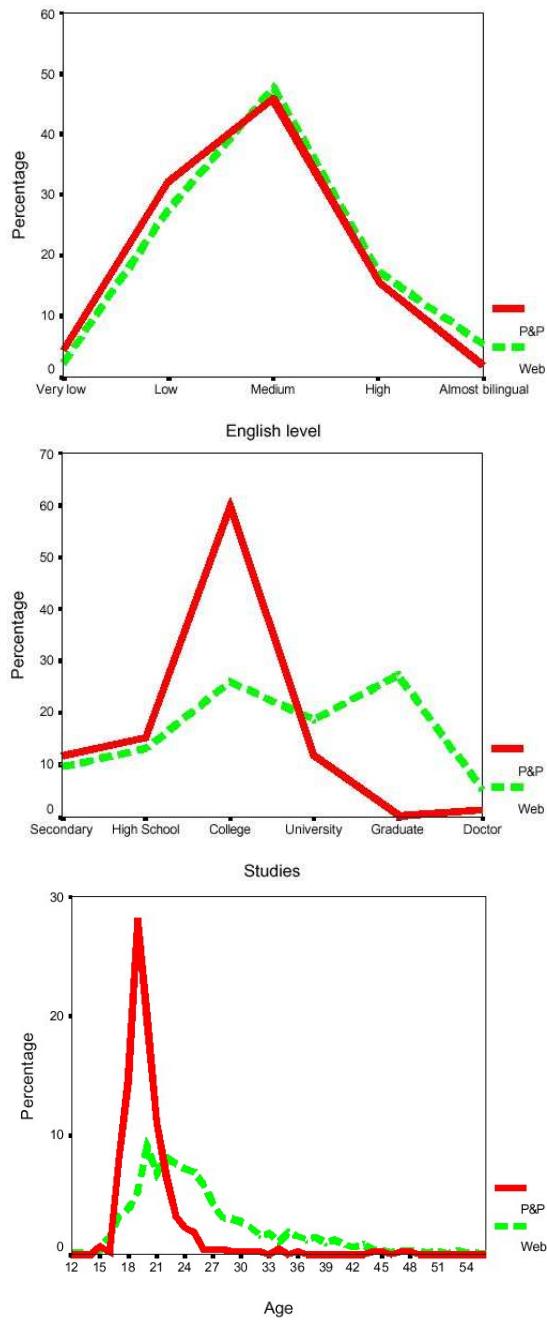


Figura 7.17: Comparativa de los perfiles de alumnos en la muestra poblacional según: nivel de inglés (arriba), educación (centro) y edad (abajo).

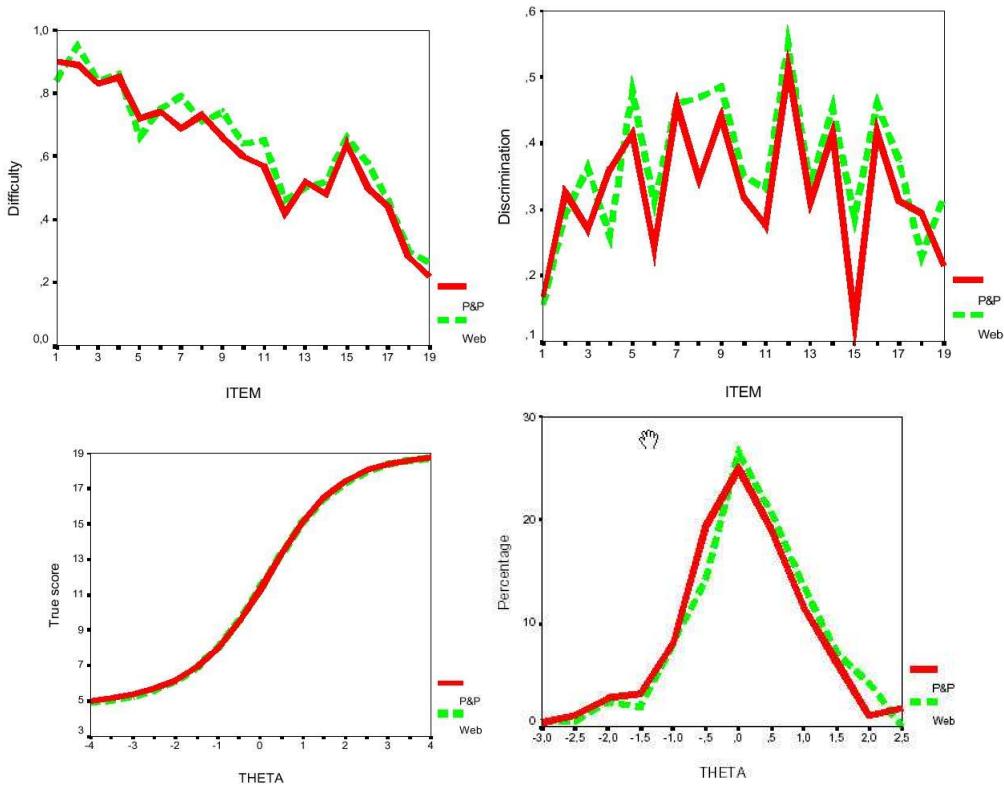


Figura 7.18: Comparación de resultados tras la calibración en función de: Arriba: dificultad (izquierda) y discriminación (derecha). Abajo: puntuación verdadera (izquierda) y niveles de conocimiento (derecha).

lizando la denomina *puntuación verdadera* del test (en inglés, *true score*). La puntuación verdadera no es más que una relación que se establece entre el nivel de conocimiento del alumno y el número de ítems del test. Se suele utilizar para expresar la calificación final del alumno en función de su puntuación verdadera, que para el alumno suele ser un dato más fácil de interpretar que el nivel de conocimiento. Para calcular la función que describe la puntuación verdadera basta con sumar todas las CCI de los ítems del test. La gráfica inferior izquierda de la figura 7.18 muestra las funciones de puntuación verdadera para ambas muestras, mostrando que ambas gráficas son muy similares.

Por último, la gráfica inferior derecha de la figura 7.18 muestra el porcentaje de alumnos que fueron evaluado con cada nivel de conocimiento, para ambas muestras. Los resultados son también similares entre las muestras, aunque no tanto como en el caso anterior. En este caso puede apreciarse que, globalmente, los niveles de conocimiento estimados de los alumnos de la muestra que realizó el test a través de SIETTE eran mayores que los obtenidos por los alumnos que hicieron el test mediante papel y lápiz.

7.8.2. Discusión

Se ha llevado a cabo un estudio de la validez de un sistema Web como medio de recolección de sesiones de test, con el objetivo final de calibrar un conjunto de ítems para su posterior administración adaptativa. En este estudio se utilizaron muestras poblacionales de un tamaño considerable. El estudio mostró diferencias existentes entre las muestras poblacionales que realizaron el test con los dos medios en las variables estudiadas: nacionalidad, sexo, edad, nivel de inglés y educación.

Por otro lado, a partir de los resultados anteriores, queda patente que la correspondencia entre las estimaciones de los parámetros de los ítems utilizando ambos medios era bastante alta. De esta forma, es posible concluir que, al menos para este test y estas muestras poblacionales, es equivalente utilizar la Web como medio de recolección de evidencias para calibración, a utilizar el sistema tradicional de administración mediante papel y lápiz.

7.9. Evaluación formativa del sistema SIETTE

La evaluación formativa mide y documenta el impacto de un programa. Su resultado final es, por tanto, un informe en el que se evalúa precisamente ese impacto. Por ejemplo, un informe de este tipo detalla quién ha participado en un programa, qué actividades se vieron afectadas, qué resultados y/o mejoras se obtuvieron de esa participación.

Winne (1993) identifica brevemente diversos propósitos que deben seguirse cuando se llevan a cabo evaluaciones de sistemas educativos. De entre ellos, los cinco siguientes pueden aplicarse claramente a la evaluación formativa:

- evaluar la eficacia de un programa de forma periódica, para así poder identificar las mejoras que son necesarias;
- verificar si las hipótesis en las que se basa el comportamiento de un programa son defendibles;
- generar información para guiar de forma más adecuada a los usuarios de un programa;
- verificar que se pone a disposición de los alumnos un programa que es realmente fiable;
- determinar, de forma más clara, las direcciones hacia las que se dirigirán las futuras versiones del programa.

En esta sección se describirá cómo se ha realizado la evaluación formativa del sistema SIETTE, durante los años en los que se ha llevado a cabo su desarrollo. El procedimiento seguido se inspira en la descripción realizada por Barros (1999), y en los pasos recomendados en (Harvey, 1998, pág. 71). El trabajo de evaluación de Barros, se fundamenta a su vez en las recomendaciones de Murray (1993) para el desarrollo de STI, y consta de los siguientes puntos:

1. *Diseño iterativo*: El diseño iterativo o cíclico está especialmente recomendado en la literatura sobre el desarrollo de sistemas que interactúan con usuarios, para asegurar la practicidad y usabilidad del software.

2. *Diseño participativo:* La evaluación continua de un sistema y su posterior refinamiento, por sí solos, no son suficientes para asegurar la usabilidad del sistema. Es necesario que los usuarios finales se vean involucrados en estos ciclos de diseño y evaluación. El objetivo es modificar la percepción que se tiene del "usuario como mero sujeto", y convertirla en "usuario como codiseñador".
3. *Evaluación formativa:* El diseño iterativo se convierte en evaluación formativa, cuando se documentan los cambios realizados sobre el sistema, las motivaciones que llevaron a esos cambios, los resultados del cambio, y en general, todos aquellos aspectos encontrados durante el estudio. La evaluación formativa puede ser cualitativa o cuantitativa. Se dice que la evaluación formativa es cualitativa cuando el objetivo es identificar las características de una determinada situación. Por el contrario, se dice que es cuantitativa cuando se intentan buscar las causas y consecuencias de esa situación.



Figura 7.19: Imagen de uno de los laboratorios de docencia de la E.T.S.I. Informática.

En general, en cada experimento que se describirá a continuación, el procedimiento que se ha seguido es básicamente el mismo. Se ha sometido a una población de estudiantes universitarios a un test a través de SIETTE. En la mayoría de los casos, la calificación del test era vinculante, es decir, formaba parte de la calificación global de los alumnos en una determinada asignatura. El test se administró en uno o más laboratorios de docencia del departamento de Lenguajes y Ciencias de la Computación de la E.T.S.I. Informática de la Universidad de Málaga. Todos estos laboratorios son bastante similares en cuanto a estructura, aunque su tamaño, y por tanto, el número de alumnos que pueden concurrir en ellos difiere. Como se puede apreciar en la figura 7.19, estos laboratorios están organizados en pares de filas de mesas, donde en cada una suele haber 8 puestos. Cada puesto tiene un PC y una silla. Los pares de filas están dispuestos de tal forma que los alumnos de una fila se sientan enfrente de los alumnos de la fila emparejada. Tanto la torre, monitor y teclado de cada puesto están situados encima de la mesa, lo que dificulta la posibilidad de que un alumno pueda comunicarse con el de enfrente. Los experimentos se han llevado a cabo siempre en laboratorios donde el sistema operativo de los PC es Windows.

Los experimentos han sido realizados, en la mayoría de las ocasiones, bajo la supervisión

de los profesores de la asignatura correspondiente, del director de esta tesis y del autor. Asimismo, en los experimentos en los que el volumen de alumnos lo ha requerido, y tal y como es norma habitual en los exámenes de las asignaturas impartidas por el departamento, personal adicional (profesores y/o becarios) han colaborado en la vigilancia de los laboratorios. De esta forma, se asegura que, como mínimo, cada laboratorio esté controlado por al menos una persona.

Inmediatamente antes de comenzar cada test, y después de que los alumnos ocuparan su puesto, se repartía a cada uno una hoja con las instrucciones, una descripción del modo de funcionamiento y características del test. Asimismo, y para evitar un uso fraudulento del sistema, los nombres de usuario y claves de los alumnos que iban a realizar el test se generaban de forma automática. De esta forma, en la hoja de instrucciones, a cada alumno se le indicaba el usuario y la clave que debían utilizar para realizar el test. Una vez que el alumno insertaba este par de valores en SIETTE, debía añadir su nombre completo (y opcionalmente su correo electrónico), el cual además debía escribir en la hoja de instrucciones. Al final del test, los alumnos estaban obligados a entregar esa hoja para que quedara constancia de que habían realizado el test.

El objetivo con el que se diseñaron estos experimentos era estudiar el comportamiento del sistema SIETTE en un entorno real; verificar que el sistema era realmente utilizable; y corregir y depurar todas aquellas incidencias que surgieran durante las pruebas.



Figura 7.20: Diagrama temporal con la evolución de los prototipos de SIETTE.

A continuación se va a proceder a detallar las características de los distintos prototipos de SIETTE, así como los experimentos que se realizaron con cada uno de ellos, y los resultados que se obtuvieron. En la figura 7.20 se muestra un diagrama temporal en el que se puede ver la evolución de los diferentes prototipos de SIETTE.

7.9.1. Primer prototipo

Aunque como se ha mencionado en el capítulo anterior, el sistema SIETTE comenzó a desarrollarse en la segunda mitad del año 2000, el primer prototipo apto para ser utilizado en un test a gran escala fue probado en el mes de febrero de 2003.

Experimentos realizados

Para probar este primer prototipo, se llevaron a cabo diversos experimentos. El primero de ellos fue realizado con estudiantes de cuarto de la titulación de Ingeniero en Informática de la Universidad de Málaga. Se trataba de un grupo formado por 85 alumnos, estudiantes

de la asignatura de *Procesadores del Lenguaje*, que es de carácter anual. El test que se realizó formaba parte de la calificación del primer cuatrimestre. La prueba se realizó el 6 de febrero de 2003, en los laboratorios de docencia de la E.T.S.I. Informática.

Era un test sobre el tema de LEX, formado por un número total de 25 ítems. Se trataba de un test heterogéneo compuesto por 9 de respuesta múltiple con un número de opciones independientes que oscilaba entre 4 y 8; 8 de respuesta corta corregidos mediante una expresión regular; y 8 de opción múltiple con un número de opciones de entre 4 y 7. Los ítems eran mostrados a los alumnos en testlets de 5 (5 por página), y disponían de un tiempo límite de 60 minutos para completar el test. El criterio de selección de ítems era aleatorio, y el de evaluación el porcentual. Además, a cada estudiante se le mostraban las opciones de respuesta de los ítems ordenadas de forma aleatoria.

El segundo experimento fue realizado, utilizando el mismo test, dos meses después (abril de 2003). Se administró a estudiantes de tercero de la titulación de Ingeniero Técnico en Informática de Gestión. Más concretamente, a los alumnos matriculados en la asignatura de *Traductores, Compiladores e Intérpretes*, de contenidos similares a la de *Procesadores del Lenguaje*. Lo realizaron un total de 101 individuos.

Resultados

En el test realizado en febrero de 2003, el alumno que lo realizó en el menor tiempo, lo hizo en 24 minutos y 25 segundos. La mayoría de los alumnos requirieron de entre 50 y 60 minutos. Todos pudieron completar el test en el tiempo establecido.

En la figura 7.21 (arriba), se muestra la distribución de alumnos, según los resultados obtenidos en el test. En el eje de abscisas se han representado los percentiles de la evaluación, y en el eje de ordenadas, el número de estudiantes cuya calificación (en porcentaje de preguntas respondidas de forma completamente correctas) era el percentil correspondiente. Como se puede apreciar, la distribución resultante se acerca bastante a una normal, centrada, en este caso, en el quinto percentil (calificaciones entre 41 % y 50 %). Asimismo, el 66 % de la muestra respondió correctamente al 50 % o más de los ítems administrados.

Es necesario destacar que en este test, a pesar de estar configurado para que todos los alumnos realizaran el mismo número de ítems, hubo un pequeño porcentaje de individuos (el 7 %) que realizó un número menor de preguntas (debido a un fallo en el sistema), tal y como se muestra en el gráfico circular situado en la parte inferior de la figura.

La figura 7.22 muestra los resultados del test administrado en abril de 2003. La gráfica superior es la distribución de los alumnos según el porcentaje de ítems respondidos correctamente. Como se puede apreciar, la distribución es también más o menos normal, aunque esta vez está centrada en el cuarto percentil. Esto permite inferir que, en general, el nivel de conocimiento de este segundo grupo de estudiantes era inferior al observado en la muestra anterior. Tan sólo el 40 % de ellos obtuvo un porcentaje de éxito mayor o igual al 50 % de los ítems administrados. Por otra parte, el gráfico circular inferior de la figura muestra el porcentaje de individuos cuya sesión de test estuvo compuesta por 25 ítems, frente al porcentaje que resolvieron un número menor de preguntas. En este caso es significativo que el porcentaje de alumnos que realizaron un número menor de ítems es bastante mayor. Además, mientras que en el primer experimento la sesión de menor tamaño era de 23 ítems, en este segundo la de menor número de preguntas estaba compuesta tan sólo de 14. En cuanto al alumno que necesitó menor tiempo para completar el test, éste requirió aproximadamente 39 minutos.

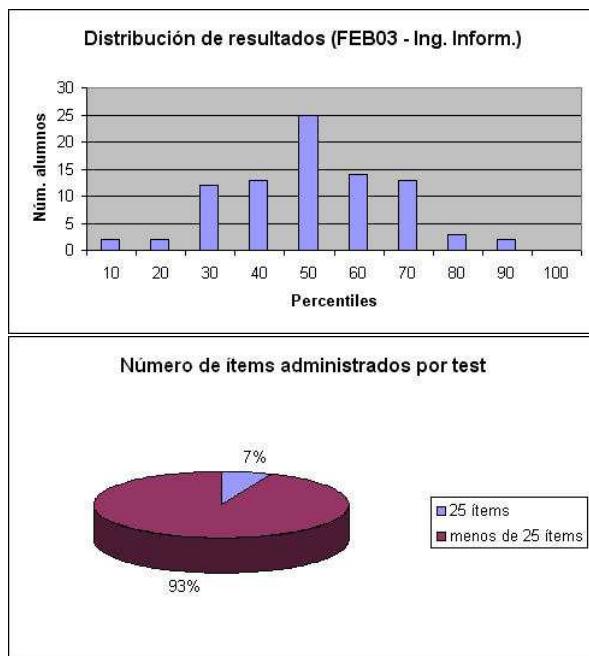


Figura 7.21: Resultados del test de Procesadores del Lenguaje administrado en febrero de 2003.

En cuanto a las incidencias que surgieron durante el transcurso de ambos experimentos, además del hecho de que, por algún motivo, no todos los alumnos realizaron el mismo número de ítems, se puede destacar las siguientes:

- Algunos examinandos tuvieron problemas al comenzar el test, debido a que el navegador del PC en el que estaban, no tenía instalado adecuadamente el *plugin* para soportar applets. Este *plugin* es esencial en tests temporizados, puesto que el cronómetro que controla el tiempo es un applet.

El principal inconveniente es que los alumnos sólo son conscientes de este problema en el momento que intentan enviar el primer conjunto de respuestas, ya que el sistema no permite esta operación, lo que les obliga a cambiar de PC y a comenzar el test de nuevo.

- Debido a problemas de conexión a Internet (ajenos al sistema), algunos estudiantes perdieron temporalmente la conexión con SIETTE, lo que les obligó a tener que volver a comenzar el test desde el principio.
- El tiempo que transcurre entre la fase de autenticación del alumno en el sistema, y el momento en el que se le muestra el primer conjunto de ítems es excesivamente largo.
- El tiempo que transcurre entre que un alumno envía las respuestas a un conjunto de ítems hasta que se le muestra el siguiente grupo es también excesivamente largo.

En cuanto a la opinión general de los alumnos, éstos manifestaron las siguientes inquietudes:

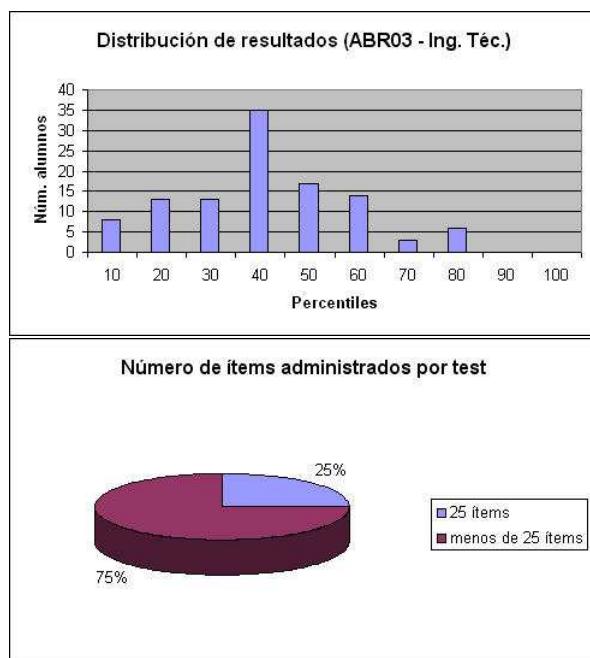


Figura 7.22: Resultados del test de Compiladores, Traductores e Intérpretes administrado en abril de 2003.

- La pregunta que más frecuentemente realizaron a los profesores era si el intervalo que transcurría entre el envío de las respuestas y la recepción de los siguientes ítems, contabilizaba en el tiempo total del que disponían para completar el test. A los alumnos se les explicó que el reloj se detiene en el momento de envío, y que se vuelve a activar cuando se les muestra el siguiente conjunto de ítems. Aún así, se podía apreciar que el excesivo retardo de tiempo fue un factor que estresó considerablemente a los alumnos, lo cual pudo afectar a su rendimiento en el test.
- Otra queja bastante común fue el hecho de que, en los ítems de opción múltiple, una vez que se selecciona una de las opciones, no se puede dejar el ítem en blanco (deseleccionar la opción previamente elegida). Se les explicó que se trataba de un problema propio de Javascript (lenguaje que controla los componentes de los formularios de las páginas Web).
- Los estudiantes manifestaron que la forma de evaluar los ítems de respuesta múltiple con opciones independientes era muy injusta; puesto que una pregunta de este tipo se evalúa como correcta si lo está completamente, y en otro caso se considera incorrecta, es decir, no permite considerar correcciones parciales.
- Los alumnos se mostraban también preocupados por el mecanismo de corrección de los ítems de respuesta corta. Se les explicó que éste no se limitaba a comparar su respuesta con la correcta, sino que se utilizaba una expresión regular, la cual permitía un abanico más amplio de posibilidades.
- Finalmente, otra queja de los estudiantes era la que SIETTE no permitiera modificar la respuesta a un ítem una vez que ésta había sido enviada.

El aspecto más valorado positivamente es el hecho de que las calificaciones obtenidas se muestran inmediatamente una vez que el test ha finalizado. Este aspecto ha sido puesto de manifiesto frecuentemente tanto por profesores como por alumnos.

7.9.2. Segundo prototipo

Los problemas que surgieron durante la evaluación del prototipo anterior dieron lugar al segundo. Este último fue probado en la convocatoria ordinaria de exámenes de septiembre de 2003. Las características que añade este nuevo prototipo se enumeran a continuación:

- Una vez que se autentifica el alumno, en la página de descripción del test, para aquellos tests que son temporizados o contienen applets, se muestra un applet de ejemplo, y se advierte a los alumnos que, para poder realizar el test sin ningún tipo de problema, deben ver correctamente el applet.
- Se añade la posibilidad de que los tests puedan interrumpirse y retomarse de nuevo con posterioridad. Esta característica permite que si mientras que un alumno está realizando un test, se pierde la conexión con SIETTE, éste pueda continuarlo, desde otro PC, a partir de donde lo dejó, sin necesidad de comenzar de nuevo.
- Se estudió el tiempo empleado en cada una de las acciones que SIETTE lleva a cabo desde el momento que el alumno envía la respuesta a un ítem, hasta que se le muestra el siguiente. Se descubrió que el cuello de botella estaba en las instrucciones que actuallizaban la estimación del conocimiento del alumno en su modelo. Se decidió eliminar este paso, y almacenar únicamente la estimación tras la finalización del test. A pesar de que no se almacenen las inferencias parciales, si se produjera una interrupción en una de las sesiones, estos valores son fácilmente inferibles (indirectamente) a partir de los ítems administrados y sus respuestas; información que sí se almacena en modelo del alumno periódicamente.
- Se estudiaron también las razones que podían haber provocado que algunas sesiones de los tests administrados con el prototipo anterior tuvieran un número de ítems menor de 25 (tamaño establecido en el test), aún cuando no había expirado el tiempo máximo. Se apuntó la posibilidad de que fuera debido a que algunos individuos, por el excesivo intervalo que transcurría entre el envío de las respuestas y la presentación de los siguientes ítems, pulsaran repetidas veces sobre el botón de enviar. Para verificar esta hipótesis, en este prototipo, cada vez que el alumno pulsaba el botón de enviar, éste se deshabilitaba automáticamente.
- Se añadió también la posibilidad de establecer *cadencias* (tiempo mínimo entre dos ejecuciones de un mismo test) a nivel de usuario y de IP que evitaran que un determinado alumno pudiera repetir un mismo test hasta que no transcurriera un intervalo de tiempo, indicado por el profesor en el test. La cadencia por IP es un concepto análogo al anterior, según el cual no se puede repetir un mismo test. En este caso en vez de controlarse por usuario, se controla según la dirección IP desde la que se conecta ese usuario.
- En las instrucciones en papel del test, se les indica a los alumnos, de forma explícita, que en los ítems de opción múltiple, una vez que se selecciona una opción el ítem no puede dejarse en blanco; sólo puede modificarse la opción seleccionada.

Experimentos realizados

Utilizando este prototipo se administraron dos tests el 9 de septiembre de 2003, uno a continuación del otro, como parte de la evaluación de la asignatura de *Procesadores del Lenguaje*. El primero de ellos era similar a los administrados sobre el primer prototipo, es decir, se trataba de un test sobre el tema de LEX, en este caso, con un número total de 18 ítems. Los alumnos disponían de un tiempo máximo de 45 minutos para responder a todas las preguntas. Al igual que en las dos experiencias anteriores, los ítems se elegían utilizando el criterio de selección aleatoria, y la evaluación era porcentual.

El segundo test evaluaba a los estudiantes sobre análisis sintáctico, más concretamente sobre los conceptos de *cabecera* y *siguiente*. Tenía una tamaño de 15 ítems, y disponían de un tiempo máximo de 30 minutos para completarlo. Todos los ítems del test eran de respuesta múltiple con un número total de 9 opciones de respuesta independientes. Éstos se mostraban siguiendo el criterio de selección aleatorio y en testlets de 5 ítems. El mecanismo de evaluación, al igual que en el test anterior, era el porcentual.

Resultados

La figura 7.23 muestra las distribuciones correspondientes a los tests administrados en septiembre de 2003. Como se puede apreciar, con respecto al test administrado en febrero del mismo año, el de *LEX*, o bien contenía ítems menos difíciles, o bien el nivel de conocimiento de los alumnos que lo realizaron era considerablemente mayor (el 80 % de los individuos acertó un porcentaje mayor o igual al 50 % de los ítems administrados). En este caso, el diagrama de barras con la distribución de calificaciones se aleja de una normal. También es cierto que se trata de una muestra poblacional de un tamaño bastante reducido, tan sólo 23 alumnos.

En cuanto al test sobre los conceptos de *Cabecera* y *Siguiente*, la distribución de calificaciones tampoco se asemeja a una normal. Además, en este caso, el porcentaje de acierto era considerablemente mayor: un 87 % de los individuos acertó un porcentaje de ítems mayor o igual al 50 %, y de hecho el percentil más frecuente fue el décimo (seis de los 23 alumnos obtuvieron un porcentaje entre el 91 % y el 100 %).

En lo referente a las incidencias observadas durante la administración del test, tanto en uno como el otro, en todas las sesiones realizadas, el número de ítems administrado a los alumnos correspondía al establecido en el test. Esto parecía confirmar la hipótesis de que el número menor de preguntas en algunas sesiones era debido al hecho de que algunos estudiantes pulsaban varias veces el botón de enviar respuestas, lo que provocaba inconsistencias en las sesiones.

Asimismo, el tiempo de espera entre el envío por parte del alumno de las respuestas a un conjunto de ítems, y la administración del siguiente conjunto se reduce notablemente. Por este motivo, en esta prueba no hubo quejas de los examinandos en referencia a este aspecto. Tan sólo en lo referente al tiempo que transcurre desde que se autentificaban hasta que se mostraba el primer lote de ítems del test.

Aunque se recordó a los alumnos que debían verificar, en la página Web que se muestra inmediatamente antes del comienzo del test, que su navegador estaba correctamente configurado para visualizar applets; un número reducido de ellos no lo hizo, y por lo tanto tuvo que cambiar de puesto poco tiempo después de haber comenzado el test. Otro grupo de alumnos tuvo problemas de conectividad con el sistema, lo que provocó que tuvieran que

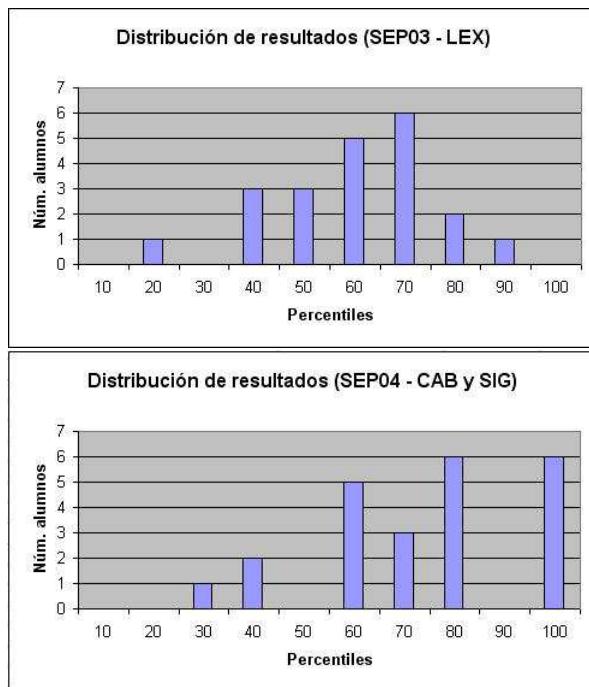


Figura 7.23: Resultados de los tests de Procesadores del Lenguaje administrados en septiembre de 2003.

reiniciar el test. Esto permitió verificar que el procedimiento implementado que permite recuperar el estado de una sesión interrumpida, funcionaba correctamente.

En cuanto a las quejas de los alumnos, además de la previamente mencionada sobre el tiempo de carga del banco de ítems, volvieron a repetirse las mismas sobre la evaluación de los ítems de respuesta múltiple con opciones independientes, y la imposibilidad de modificar las respuestas una vez enviadas.

7.9.3. Tercer prototipo

Este tercer prototipo surge (principalmente) de las modificaciones realizadas para adecuar SIETTE a los requisitos de un test sobre LISP administrado en diciembre de 2003. Entre las características que añade este nuevo prototipo, se pueden destacar las siguientes:

- Se incluye un nuevo mecanismo de evaluación heurístico, el criterio por puntos. Éste fue explicado detalladamente en capítulos anteriores. Permite calificar parcialmente los ítems de respuesta múltiple con opciones independientes, así como configurar los tests para penalizar las respuestas incorrectas.
- Se diseña un nuevo procedimiento para la carga inicial del banco de ítems del test. Éste sólo se lleva a cabo una vez de forma global para todos los alumnos que van a realizarlo. Además, se construye una caché software que permite a los profesores precargar el banco de ítems del test antes de que éste dé comienzo. Esto se traduce

en que el tiempo desde que el examinando se autentifica en SIETTE hasta que se le muestra la primera pregunta se reduce notablemente.

- Se añade la posibilidad de crear grupos de alumnos, y restringir el acceso a los tests a únicamente ese conjunto de individuos. Esta característica evita que examinandos que no estén realizando el test desde los laboratorios de docencia puedan tener acceso al mismo desde un PC externo.
- Por último, se diseña un mecanismo que permite restringir el número de ocurrencias de un ítem generativo en una misma sesión de test. Este nuevo valor se traduce en un parámetro adicional asociado al ítem generativo. Asimismo, en los criterios de selección se priorizan otros ítems, para evitar que una misma pregunta de este tipo sea elegida inmediatamente después de haber sido administrada.

Experimentos realizados

El primer test que fue administrado con este nuevo prototipo era sobre el lenguaje de programación declarativo LISP, y evaluaba el conocimiento que de éste tenían los alumnos. Esta prueba de evaluación se llevaba realizando desde el curso 1990/91 en formato de papel y lápiz, y a partir de diciembre de 2003 se administra utilizando el sistema SIETTE. Es parte de la asignatura de *Inteligencia Artificial e Ingeniería del Conocimiento* (de carácter anual), que se imparte en cuarto curso de la titulación de Ingeniero en Informática, y que también pertenece a la docencia asociada al departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga.

Este test de LISP constaba de 20 ítems de opción múltiple, todos ellos con tres opciones de respuesta. Éstos eran mostrados de 5 en 5, y el alumno disponía de un tiempo máximo de 25 minutos para finalizarlo. El criterio de selección era el aleatorio, y el de evaluación era el nuevo mecanismo por puntos. Cada respuesta correcta sumaba un punto, mientras que una respuesta incorrectamente seleccionada restaba 0,5 puntos.

Como característica adicional hay que reseñar que este test fue el primero elaborado por profesores que no había utilizado ni tenían conocimientos previos del funcionamiento de SIETTE. A éstos se les explicó brevemente cómo se creaba el currículo, ítems y tests de una asignatura. Con esta información pudieron construir el test sin incidencias notables.

Posteriormente, en ese mismo mes de diciembre, se administró otro test de Procesadores del Lenguaje sobre LEX, compuesto por un número total de 20 preguntas, en el que se combinaban ítems de opción múltiple, de respuesta múltiple con opciones independientes y de respuesta corta. Esta vez el criterio de evaluación era el *por puntos*, pero sin penalización por respuesta incorrecta.

En febrero de 2004, se realizó un test de una nueva asignatura: *Laboratorio de Ingeniería del Software*. Se trata de una optativa de segundo ciclo, de la titulación de Ingeniero de Telecomunicaciones, eminentemente práctica, en la que los alumnos, por grupos, deben realizar una práctica en el lenguaje de programación Java. Asimismo, y como novedad a partir de febrero de 2004, deben realizar un test a través de SIETTE, que evalúa sus conocimiento sobre el lenguaje Java, y que se tiene en cuenta como parte de la calificación final en la asignatura.

El test estaba formado por 50 ítems en los que se combinaban de opción múltiple, de respuesta múltiple con opciones independientes y de respuesta corta. Éstos se mostraban de 5 en 5, y el alumno disponía de un tiempo total de dos horas para finalizar el test. El

criterio de selección era el aleatorio y el mecanismo de evaluación el heurístico por puntos con penalización por respuesta incorrecta.

También en febrero de 2004 se administraron dos tests (uno detrás de otro) de la asignatura de Procesadores del Lenguaje. El primero de ellos evaluaba a los alumnos sobre *análisis SLR(1)*, y se componía de 15 ítems: 9 de respuesta múltiple, con 18 opciones independientes cada uno de ellos; 3 de opción múltiple con 4 opciones; y 3 de respuesta corta. Las preguntas eran mostradas de 3 en 3, y los examinados tenía un tiempo total de 45 minutos para completarlas. El criterio de selección empleado fue el ordenado, y el método de evaluación el heurístico porcentual.

El segundo de los tests evaluaba el conocimiento de los alumnos sobre LEX, y sólo estaba formado por un único ítem de respuesta múltiple con 25 opciones independientes. La idea era emular los testlets en los que todas las preguntas que los integran tienen un común un enunciado global. En este caso, el único ítem podía considerarse como un testlet formado por preguntas verdadero/falso. Los examinados disponían de 15 minutos para resolver este test. El criterio de evaluación utilizado fue el heurístico por puntos, donde el ítem tenía asignado 25 puntos (uno por opción correctamente seleccionada).

En junio de 2004 se realizaron tres tests sobre el lenguaje Java dentro de la asignatura de *Laboratorio de Tecnología de Objetos* que se imparte en las titulaciones de Ingeniero en Informática, Ingeniero Técnico en Informática de Sistemas e Ingeniero Técnico en Informática de Gestión de la Universidad de Málaga, durante el segundo cuatrimestre del segundo curso de las tres titulaciones. Los tres tests eran bastante similares y se componían de 10 ítems de opción múltiple, todos ellos con cuatro opciones de respuesta. Los estudiantes disponían de 20 minutos para completar el test, y los criterios de selección y evaluación eran el aleatorio, y el por puntos, respectivamente. En este caso, cada pregunta acertada suponía un punto en la calificación final, mientras que una respuesta incorrecta restaba 0,33 puntos.

Estos últimos tests también fueron elaborado por profesores sin conocimientos previos del funcionamiento de SIETTE. Al igual que en el caso anterior, a éstos se les explicó brevemente el procedimiento para la construcción de los elementos de una asignatura, a través de la herramienta de autor; y también elaboraron los tests sin incidencias significativas.

Resultados

La figura 7.24 muestra los resultados del test de LISP que fue administrado el 18 de diciembre de 2003. La gráfica superior muestra las frecuencias de los resultados obtenidos por los alumnos, según el criterio por puntos. En la gráfica inferior, se han comparado estos resultados, con los que se hubiesen obtenido en caso de que el mecanismo utilizado hubiese sido el porcentual. En el diagrama de barras que está delante se ha vuelto a representar la frecuencia de resultados según el criterio por puntos (con penalización), mientras que en el de detrás se ha representado la frecuencia de resultados según el porcentual.

Este test fue realizado por un total de 83 alumnos, de los cuales el 66 % obtuvo un aprobado (calificación sobre 100 mayor o igual que 50). Por el contrario, si se hubiera aplicado el criterio porcentual, esta cifra hubiera sido considerablemente mayor (un 89 % de la muestra).

En cuanto a las incidencias más destacadas, hay que mencionar que este test requirió de un control más exhaustivo por parte de los profesores, puesto que en los PC donde se administraron los tests estaba instalado un programa intérprete de LISP que los alumnos conocían y utilizaban habitualmente durante el transcurso del curso. Dado que en la mayoría

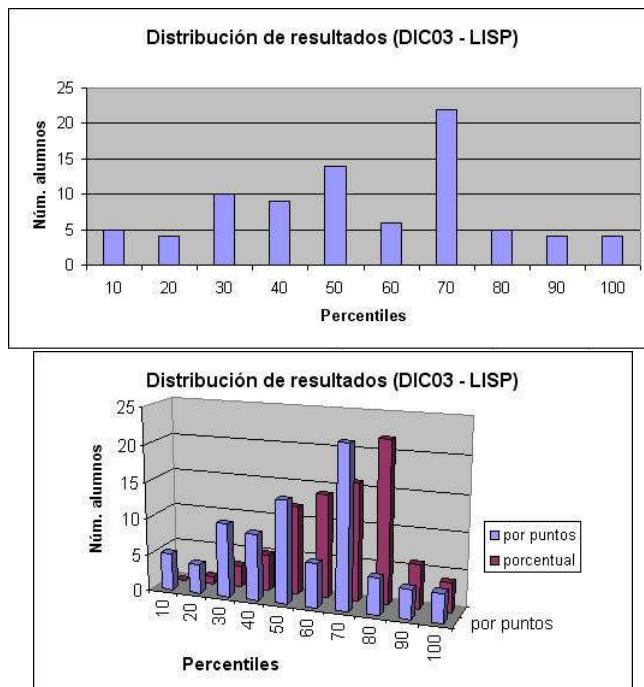


Figura 7.24: Resultados del test de LISP administrado en diciembre de 2003.

de los ítems se preguntaba a los examinandos sobre el resultado de evaluar una determinada expresión en LISP, era especialmente necesario controlar que no hiciesen uso indebido de esta herramienta. De hecho, fueron detectados un número reducido de casos en los que ciertos estudiantes tenían la aplicación con el intérprete abierto. Por esta razón, los profesores que vigilaban el examen debían verificar frecuentemente que todos los alumnos tenían abierto únicamente el navegador Web en sus PC.

Dos tests de las mismas características (aunque con diferentes ítems) fueron posteriormente administrados el 17 de febrero y el 16 de junio de 2004. En el primero de ellos, 44 alumnos realizaron el test, mientras que en el segundo tan sólo 9. Estos dos tests representaban oportunidades adicionales que los profesores de la asignatura daban a los estudiantes para superar el test. Por este motivo, sólo se presentaron a ellos aquellos alumnos que no hubieran (o no hubieran realizado todavía) el test de LISP, o aquéllos que desearan mejorar su calificación.

Por otra parte, la figura 7.25 muestra los resultados del test de Procesadores del Lenguaje. Éste fue realizado por un total de 83 alumnos. En la gráfica (diagrama de barras inferior), se muestra la comparación de resultados entre los criterios de evaluación por puntos (diagrama de barras de delante) y el porcentual (diagrama de barras de atrás). En este caso, los alumnos se ven claramente beneficiados con la evaluación por puntos. Esto es normal, puesto que sí se tienen en cuenta las respuestas parcialmente correctas a los ítems de respuesta múltiple con opciones independientes. Mientras que según el criterio porcentual sólo aprueba el 75 % de los alumnos, utilizando evaluación por puntos, este porcentaje se ve incrementado hasta el 83 %.

La figura 7.26 muestra los resultados del test de Laboratorio de Ingeniería del Software

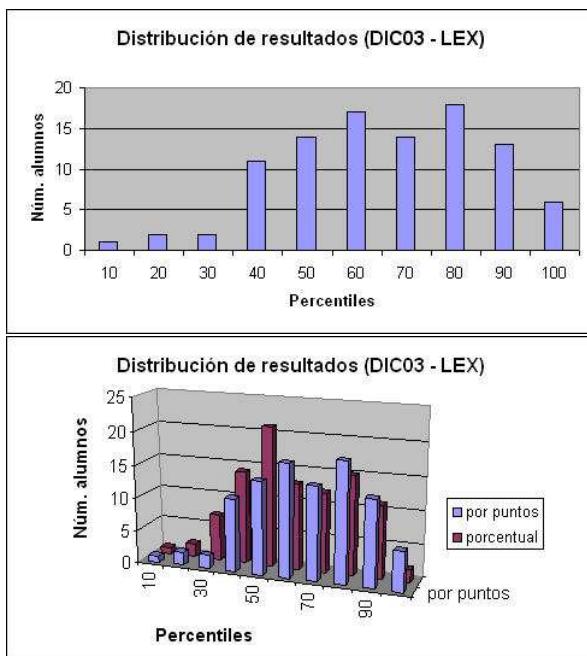


Figura 7.25: Resultados del test de Procesadores del Lenguaje administrado en diciembre de 2003.

administrado el 10 de febrero de 2004. Éste fue realizado por un total de 84 alumnos. Como se puede apreciar en la figura, la distribución de los resultados es más o menos normal centrada en el quinto percentil. El 82 % de los examinandos superó esta prueba con una calificación mayor o igual a 50 (sobre un máximo de 100 puntos).

La figura 7.27 muestra los resultados de los tests de LEX y gramáticas SLR(1) que fueron administrados en febrero de 2004. Ambos fueron realizados por un total de 105 alumnos. Como se puede apreciar en el diagrama de barras superior, la distribución de los resultados, en este caso, era bastante dispersa. El valor máximo se encuentra simultáneamente en el segundo percentil y en el sexto. Además, tan sólo el 55 % de los individuos aprobó este test. Por el contrario, los resultados obtenidos en la prueba SLR(1) son notablemente mejores (diagrama de barras inferior). En este caso, la distribución tiene un máximo diferenciado en el sexto percentil, y además, el 80 % de los examinandos superó el test.

Finalmente, en las gráficas de la figura 7.28 se han representado los resultados correspondientes a los tests de Java para la asignatura de Laboratorio de Tecnología de Objetos. La primera de ellas (figura 7.28 (superior)), muestra la distribución por percentiles correspondiente a los alumnos de la titulación de Ingeniero en Informática, cuyo máximo valor se localiza en el cuarto. El test fue realizado por un total de 79 estudiantes, de los cuales el 40 % obtuvo una calificación superior a 5.

El diagrama de barras intermedio (figura 7.28 (medio)) corresponde a los estudiantes matriculados en la titulación de Ingeniero Técnico en Informática de Gestión. Este test fue realizado por 89 individuos, de los cuales el 37 % obtuvo un aprobado. Al igual que en la distribución que expresa los resultados del test de la titulación superior, el máximo corresponde al cuarto percentil.

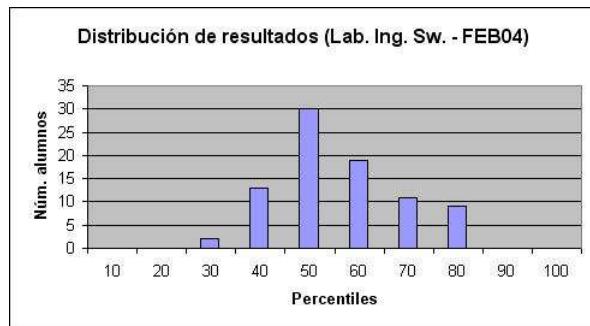


Figura 7.26: Resultados del test de Laboratorio de Ingeniería del Software administrado en febrero de 2004.

El diagrama de barras de la figura 7.28 representa los resultados de los alumnos matriculados en la titulación de Ingeniero Técnico en Informática de Sistemas. Este test fue administrado a un total de 107 individuos, de los que tan sólo un 24 % obtuvo un aprobado. A diferencia de las dos distribuciones anteriores, el valor máximo corresponde al tercer percentil.

Por último, hay que mencionar también que en este prototipo se detectó un fallo según el cual, una vez que el alumno se autentificaba, si el nombre de usuario y/o contraseña habían sido mal introducidos, en vez de mostrarse el error correspondiente, la aplicación daba uno de ejecución. Se trataba de un fallo menor, puesto que no repercutía en la ejecución normal. Para solventar el problema el examinando únicamente debía autentificarse correctamente.

7.9.4. Cuarto prototipo

El último prototipo desarrollado hasta el momento, fue probado por primera vez en septiembre de 2004. En cuanto a las características que éste añade, se pueden destacar las siguientes:

- Se corrige el error detectado en el prototipo anterior, el cual se producía cuando un alumno se autentificaba de forma errónea.
- En los ítems de opción múltiple se añade un botón que permite deseleccionar la opción de respuesta elegida por el alumno, permitiendo de esta forma, dejar la pregunta en blanco.
- En los ítems de respuesta corta, se añade la posibilidad de incluir diversos tipos de patrones que permiten capturar el tipo de respuesta. Éstos pueden ser de respuestas correctas o bien de respuestas incorrectas.
- Para que SIETTE sea soportado incluso por aquellos navegadores que tengan deshabilitadas las *cookies*, se implementa un mecanismo de paso de parámetros entre las páginas de la aplicación, a través del procedimiento denominado *reescritura de URL*.

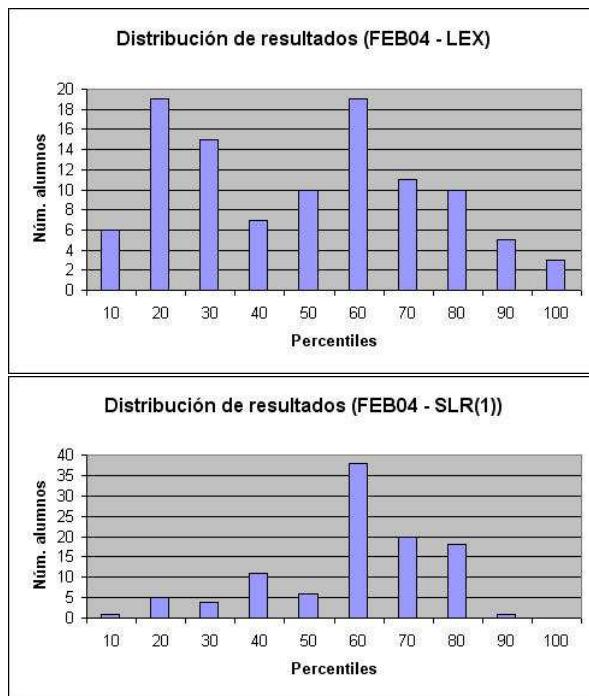


Figura 7.27: Resultados de los tests de Procesadores del Lenguaje administrados en febrero de 2004.

Experimentos realizados

El primer test que se administró utilizando este prototipo fue un test de LISP. Éste fue realizado el 9 de septiembre de 2004. El número máximo de ítems era igual a 20, y los alumnos disponían de un tiempo total de 25 minutos para completarlo. El criterio de selección era el aleatorio, y la evaluación se realizó según el mecanismo por puntos con penalización por respuesta incorrecta. Todos los ítems del test eran de opción múltiple con tres opciones de respuesta.

El 21 de septiembre del mismo año, se administró un test de la asignatura *Procesadores del Lenguaje*, correspondiente al curso 2003/04. Éste se componía de un total de 30 ítems, presentados de cinco en cinco. Todos ellos eran de opción múltiple con cuatro posibles respuestas. Los alumnos disponían de un tiempo límite de 45 minutos para completarlo. El criterio de selección fue el aleatorio, y el de evaluación el por puntos con penalización por fallo. Cada pregunta respondida correctamente sumaba un punto, mientras que si se hacía incorrectamente restaba 0,33. Este test a diferencia de los realizados hasta ahora cubría los contenidos relativos al segundo cuatrimestre de la asignatura de Procesadores del Lenguaje.

El 16 de diciembre de 2004 se administró otro test de la asignatura de Procesadores del Lenguaje, correspondiente a la parte del LEX de curso 2004/2005. Éste se componía de 20 ítems mostrados de uno en uno. Éstos eran de opción múltiple (con entre 4 y 7 opciones de respuesta), de respuesta múltiple con opciones independientes (con entre 4 y 8), e ítems de respuesta corta. En este test se incluyó la novedad de que cada ítem ofrecía al alumno un conjunto de ayudas. Éstos disponían de un tiempo máximo de 45 minutos para completarlo.

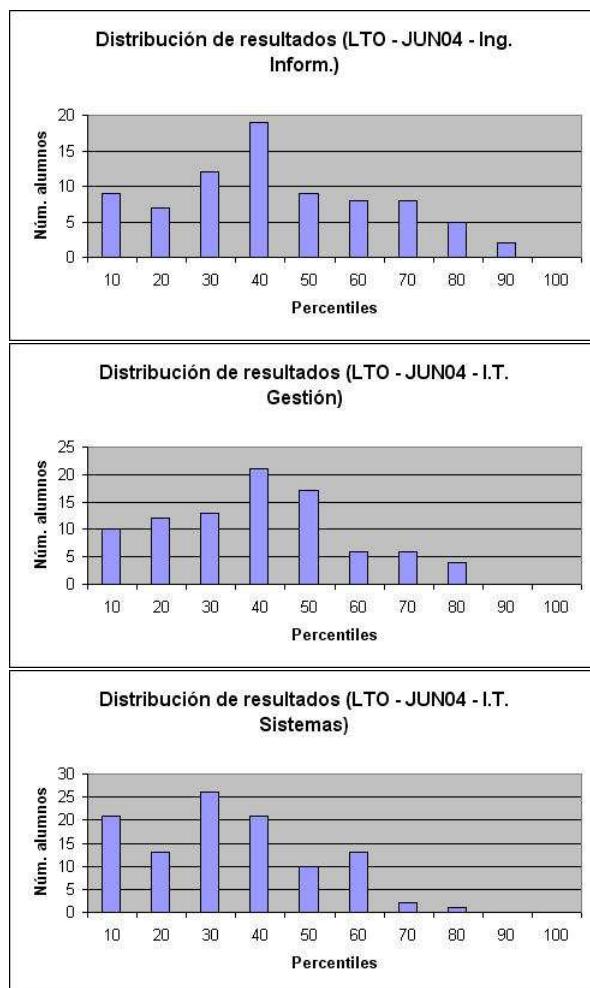


Figura 7.28: Resultados de los tests de Laboratorio de Tecnología de Objetos administrados en junio de 2004.

El criterio de selección era el aleatorio, y el de evaluación era el por puntos con penalización por el uso de ayudas. Si el alumno requería del uso de una de ellas para resolver un ítem, su puntuación (en esa pregunta) se reducía en un 75 %. Igualmente se penalizaba también por respuesta incorrecta. El objetivo de este test era analizar cómo afecta el uso de ayudas en la respuesta del alumno. La idea principal que subyace de este estudio es que, además de las curvas características que se asocian a un determinado ítem, si éste permite el uso de ayudas, es necesario estimar una nueva curva característica del ítem que esté condicionada por el uso de esa ayuda. De esta forma, tras calibrar las curvas del ítem y de éste combinado con cada ayuda, es posible determinar la utilidad de estas ayudas, puesto que si los valores de las probabilidades de la CCI de una de ellas son menores que los de la curva sin ésta quiere decir que esa ayuda no es adecuada, puesto que no contribuye a que el alumno responda correctamente al ítem (Conejo et al., 2005). Tal y como se pondrá de manifiesto en el capítulo siguiente, en la actualidad se está trabajando en esta línea y se espera, por

tanto, obtener más resultados en un futuro próximo.

El 21 de diciembre del 2004 se administró el test de LISP correspondiente a los alumnos del curso 2004/05. Como es habitual, éste estaba compuesto por 20 ítems de opción múltiple con cuatro posibles opciones de respuesta. El criterio de selección era el aleatorio, y el de evaluación el por puntos con penalización por respuesta incorrecta. Los examinandos disponían de un tiempo límite de 25 minutos para completar la prueba.

El 25 de enero de 2005 se repitió un test de las mismas características que el anterior (pero con diferentes ítems) para aquellos alumnos que no hubiesen aprobado el test de diciembre, para los que querían subir nota, o bien para aquéllos que no se hubiesen presentado al anterior.

El 3 de febrero de 2005 se administraron dos tests (uno tras otro) de la asignatura de Procesadores del Lenguaje. El primero de ellos era sobre LEX, y constaba de tan sólo 5 ítems de respuesta corta. Éste tenía la particularidad de que fue corregido a posteriori. A partir de las respuesta de los alumnos, se extrajeron los patrones de respuesta más frecuentes, los cuales a su vez se incluyeron como patrones de respuesta del ítem. Los ítems iban apareciendo de uno en uno, y cada examinando disponía de un tiempo máximo de 20 minutos para finalizarlo.

El segundo test se centraba en las *gramáticas LL(1)*. Constaba de 21 ítems, y los alumnos disponían de un total de 60 minutos para completarlo. Las preguntas se iban administrando de tres en tres, y el criterio de selección en este caso fue el ordenado. El mecanismo de evaluación fue el porcentual. Los ítems eran de respuesta múltiple con opciones independientes (con entre 4 y 12 opciones de respuesta) y de respuesta corta con entre 1 y 3 respuestas.



Figura 7.29: Imagen de uno de los laboratorios de docencia de la Escuela Técnica de Ingenieros Forestales de la Universidad Politécnica de Madrid.

Finalmente, en mayo de 2005 se llevó a cabo un test, pero esta vez no fue administrado en la E.T.S.I. Informática de la Universidad de Málaga. El objetivo era probar SIETTE en un

entorno en el que la mayor parte de los examinandos no tuvieran amplios conocimientos de informática. La prueba que se llevó a cabo estaba formada por 100 ítems de opción múltiple (con cuatro opciones de respuesta), de respuesta múltiple con opciones independientes (con cuatro opciones), y de respuesta corta corregidas mediante una expresión regular. El criterio de finalización era el basado en el número máximo de ítems, el mecanismo de selección de ítems el aleatorio, y la evaluación se realizó siguiendo el método porcentual. Este test evaluaba sobre la materia impartida en la asignatura de botánica que se imparte en segundo curso de la titulación de Ingeniero Técnico Forestal de la Universidad Politécnica de Madrid. Los alumnos realizaron esta prueba desde un laboratorio de esa Universidad (figura 7.29), sobre la versión de SIETTE instalada en Málaga. El test era de carácter voluntario y se llevó a cabo siguiendo la dinámica habitual. A cada examinando se le dio una hoja con las instrucciones, junto con el nombre de usuario y la contraseña que debían utilizar. Los ítems fueron creados por el profesor de la asignatura, que ha construido un banco con 743 ítems.

Una vez finalizado el test, los alumnos llenaron un cuestionario (figura 7.30) acerca de su opinión sobre el sistema. Éste estaba formado por 32 preguntas organizadas en cinco bloques. El primer bloque (las cinco primeras) se centraba en su experiencia en sistemas informáticos en general, y en sistemas de enseñanza virtuales. El objetivo del segundo bloque (tres preguntas) era que el examinando valorara la actividad en sí. El tercero estaba formado por 19 preguntas sobre SIETTE. El cuarto contenía una única pregunta en la que el estudiante debía valorar la experiencia global de utilizar este sistema. Las respuestas a las cuestiones de los cuatro bloques anteriores estaban ordenadas en una escala sumativa entre 1 y 5. Por último, en el bloque quinto, formado por tres preguntas, se pedía la opinión general del alumno sobre el sistema y sobre las mejoras que debían tenerse en cuenta en posteriores versiones del mismo. Estas cuestiones eran abiertas, por lo que el alumno podía extenderse en su respuesta tanto como quisiera. Finalmente, se le dejaba abierta la oportunidad de hacer cualquier otro tipo de comentario que considerara oportuno.

Resultados

La figura 7.31 muestra los resultados del test de LISP administrado en septiembre del 2004. Éste fue realizado por tan sólo 29 alumnos. De ese conjunto de individuos el 75 % obtuvo una calificación mayor o igual a 50 puntos sobre 100.

La figura 7.32 muestra los resultados correspondientes al test de Procesadores del Lenguaje administrado en septiembre de 2004. Éste fue realizado por 39 alumnos, de los cuales, el 59 % obtuvo una calificación igual o superior a 50 puntos sobre 100.

El test de Procesadores del Lenguaje realizado en diciembre de 2004 (figura 7.33), fue administrado a un total de 80 individuos, de los cuales el 90 % lo aprobó. En cuanto al uso de las ayudas, los resultados no fueron muy esperanzadores, puesto que por ejemplo, el estudio concluyó que, para uno de los ítems, sólo una de ellas contribuía realmente a mejorar los resultados de los examinandos. Esto supone, por tanto, que el resto de ayudas no eran realmente útiles y debía ser eliminadas.

En el test de LISP realizado en diciembre del 2004, del total de 79 alumnos que lo realizaron, sólo el 63 % obtuvo un aprobado (figura 7.34 (arriba)). El de enero de 2005 fue administrado a 54 examinandos, pero el porcentaje de aprobados fue bastante mejor (81 %).

La figura 7.35 muestra los resultados de los tests de la asignatura de Procesadores del Lenguaje administrados en febrero de 2005. Ambos fueron realizados por 85 individuos. La prueba de LEX fue superada tan sólo por el 48 %, mientras que la segunda por un 80 %.

CUESTIONARIO DE UTILIZACIÓN DEL SISTEMA SIETTE					
(1 MÍNIMO, 5 MUY ALTO)					
SOBRE EL ALUMNO					
1. Mi experiencia en el manejo del ordenador es ...	1	2	3	4	5
2. Mis conocimientos de informática son					
3. Me gusta usar el ordenador..					
4. Mi experiencia con sistemas de enseñanza y/o evaluación virtuales ...					
En caso afirmativos nombrarlos					
5. Mi confianza en los sistemas de evaluación por ordenador es...					
6. En mi experiencia, creo que el uso de las nuevas tecnologías para aprender ayuda en grado...					
(1 MUY POCO, 5 MUCHO)					
SOBRE LA ACTIVIDAD					
7. Me ha gustado realizar este test de esta forma...	1	2	3	4	5
8. Me ha supuesto un nivel de esfuerzo...					
9. Me he divertido realizando esta actividad....					
(1 TOTALMENTE EN DESACUERDO, 5 TOTALMENTE DE ACUERDO)					
SOBRE EL SISTEMA SIETTE					
10. La interfaz del sistema es clara y legible. Encontré sin problema los botones/opciones cuando los necesitaba	1	2	3	4	5
11. Los colores, el tamaño de letra y la organización de los elementos de la interfaz eran claras y legibles					
12. La forma en que se presentan y secuencian las preguntas es adecuada					
13. En todo momento sabía en qué pregunta estaba y cuánto me faltaba para acabar					
14. El sistema funciona bien. He podido realizar el test sin problemas y en todo momento el sistema ha respondido como esperaba.					
15. El sistema es fácil de usar					
16. La velocidad del sistema antes de comenzar a mostrar preguntas es adecuada					
17. La velocidad con la que el sistema muestra las preguntas durante el test es adecuada					
18. Siempre entendí qué me preguntaban. La legibilidad de los enunciados de preguntas es adecuada.					
19. La forma en se rellenan las respuestas es clara y adecuada en cada tipo de pregunta					
20. La presentación de resultados es adecuada					
21. Prefiero hacer el test a través de SIETTE mejor que con test de papel					
22. Prefiero la evaluación con SIETTE a la evaluación manual del profesor					
23. En caso de haber realizado otro test de la misma materia sobre papel, ¿cómo consideras que es el test de SIETTE con respecto a los anteriores?. Si no se ha realizado dejar en blanco.					
24. Confianza en los resultados obtenidos en el test (en la forma en como se recogen y guardan los resultados) es alta					
25. ¿Te gustaría usar SIETTE en otras asignaturas?					
26. ¿Te parece útil el uso de SIETTE para estudiar y no sólo para evaluación?					
27. ¿Crees que SIETTE permite hacer actividades diferentes o más novedosas e innovadoras que en una clase convencional?					
28. ¿Consideras más ventajoso el uso de SIETTE en grupos, en vez de forma individual?					
(1 MUY MALA, 5 MUY BUENA)					
VALORACION GENERAL					
29. Valoración general sobre la experiencia de uso del sistema SIETTE	2	3	4	5	
COMENTARIOS Y MEJORAS					
30. ¿Qué cosas deberían modificarse o tenerse en cuenta para futuras versiones? ¿Qué echo en falta? ¿Qué me sobra?	<ul style="list-style-type: none"> • En la interfaz • En la forma de hacer el test • En la forma de presentar las preguntas • En la forma de formular las preguntas • En la forma de presentar los resultados • Otros comentarios 				
31. ¿Qué ventajas encuentras en el uso de SIETTE? Enumerar					
32. ¿Qué desventajas encuentras en el uso de SIETTE? Enumerar					
Incluye aquí cualquier comentario u observación que consideres oportuno					

Muchas gracias por tu cooperación

Figura 7.30: Encuesta realizada a los alumnos de botánica.

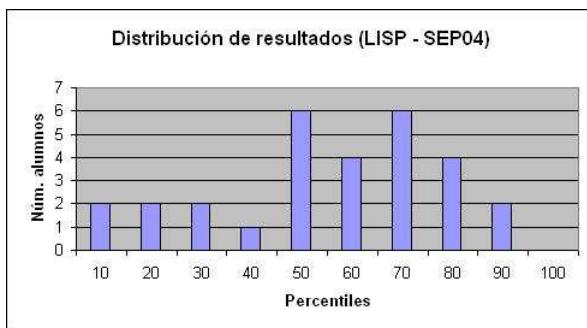


Figura 7.31: Resultados del test de LISP administrado en septiembre de 2004.

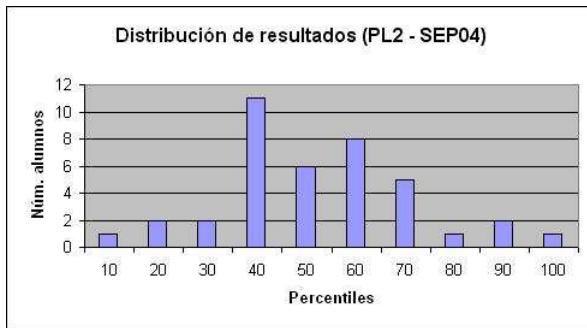


Figura 7.32: Resultados del test de Procesadores del Lenguaje administrado en septiembre de 2004.

Finalmente, en cuanto a los resultados del test de botánica administrado en mayo de 2005, éstos fueron bastante buenos; aunque también es cierto que, debido su carácter voluntario, tan sólo lo realizaron 21 estudiantes. A partir de los cuestionarios se extrajeron sus resultados, que han sido expresados en las figuras 7.36, 7.37, 7.38, 7.39, 7.40 y 7.41.

Como se puede apreciar en la figura 7.36, la mayoría de los alumnos tenían poca experiencia en el uso de PC y pocos conocimientos informáticos, en general. En cuanto a su experiencia en el uso de sistemas de enseñanza virtual, la mayoría de ellos decían tener muy poca, aunque consideraban positivo la introducción de las nuevas tecnologías en la enseñanza y evaluación. Asimismo, los estudiantes mostraban un grado de confianza medio/bajo en los sistemas de evaluación por ordenador.

La figura 7.37 muestra la valoración que de la actividad realizaron los alumnos. Se puede ver claramente que la mayoría de ellos valoraron positivamente esta experiencia y se divirtieron realizando el test. Igualmente, sobre el esfuerzo requerido para realizar el test de esta forma, según la mayoría de individuos, éste fue bastante reducido.

En el diagrama de barras de la figura 7.38 se ha representado la opinión de los alumnos sobre las características de SIETTE. La mayoría de ellos coinciden en valorar éstas de forma positiva o muy positiva. El único aspecto peor valorado fue la idoneidad de la secuenciación de las preguntas. En la sección de comentarios muchos estudiantes señalaron que hubiesen preferido que los ítems hubiesen sido presentados agrupados por tipos, en vez de mezclados.

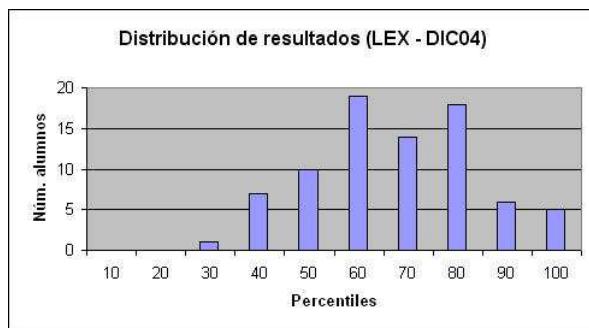


Figura 7.33: Resultados del test de Procesadores del Lenguaje administrado en diciembre de 2004.

Asimismo, también expresaron su pesar por no poder modificar las respuestas una vez enviados los resultados.

La figura 7.39 muestra los resultados del cuestionario sobre otras características de SIETTE. En cuanto a la valoración del rendimiento del sistema (su velocidad antes de comenzar el test y en la secuenciación de ítems), los alumnos valoraron muy positivamente este aspecto. Igualmente valoraron muy positivamente la presentación que hace SIETTE de las calificaciones finales del test. Otras características valoradas positivamente son la legibilidad de los enunciados y la claridad sobre cómo responder a las preguntas.

A pesar de que los estudiantes valoraron muy positivamente SIETTE, tal y como se muestra en los diagramas de barras de la figura 7.40, cuando se les preguntaba sobre si prefieren un test a través de SIETTE frente a uno de lápiz y papel, o si prefieren que éste fuera corregido y evaluado manual; aunque los resultados presentan un mayor grado de dispersión, en general, los alumnos se decantan por los métodos tradicionales. Asimismo, parecen confiar poco en los resultados del test. Igualmente, parecen bastante reacios a que SIETTE se utilice en otras asignaturas, aún cuando valoran positivamente su utilidad, y la innovación y ventajas que supone su uso.

Por último, en el diagrama circular de la figura 7.41 se han representado la valoración global que de la experiencia de uso del sistema SIETTE hicieron los estudiantes de botánica. Como se puede apreciar, ninguno de ellos valoró la experiencia de forma negativa, y de hecho la mayoría de ellos la valoraron como buena. En cuanto a las opiniones sobre las ventajas de SIETTE, los alumnos destacaron entre otras las siguientes: *la posibilidad de incluir imágenes en los enunciados y respuestas de los ítems, el hecho de conocer la calificación de forma inmediata, el test es más rápido que sobre papel, permite evaluar la capacidad de reconocer especies, el test es más ameno, más cómodo y más sencillo, etc.* En cuanto a las desventajas, los alumnos señalaron las siguientes: *algunas fotos eran pequeñas y no se aprecian con claridad, los ítems de respuesta corta no son tolerantes a fallos tipográficos, siendo por tanto más rígidos que el profesor, una vez respondido no se puede ir hacia atrás en el test, ...*

7.9.5. Conclusiones

Como se puede apreciar, las sucesivas experiencias de uso de SIETTE han permitido depurar fallos y añadir nuevas características que han mejorado la usabilidad del sistema. De

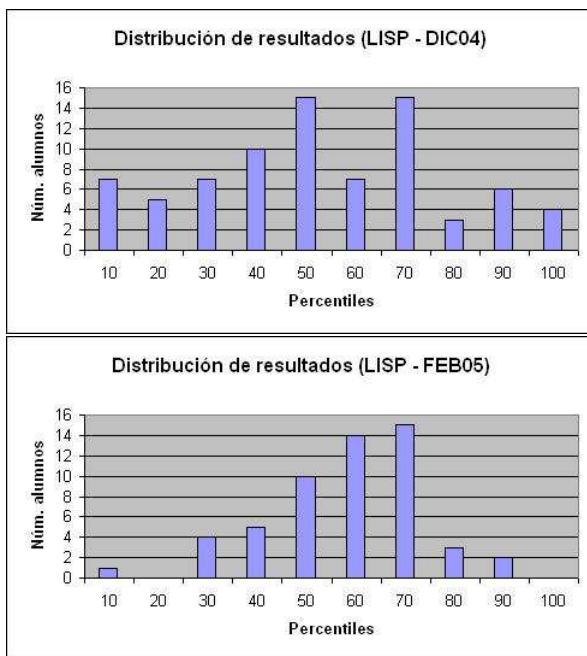


Figura 7.34: Resultados de los tests de LISP administrados en diciembre de 2004 y enero de 2005.

esta forma, gracias al proceso de evaluación formativa que se ha llevado a cabo, el sistema se ha enriquecido notable. Asimismo, el test realizado en la Universidad Politécnica de Madrid en mayo de 2005 ha puesto de manifiesto la ubicuidad real de SIETTE, concluyendo que éste apto no sólo para ser utilizado en la Universidad de Málaga, sino que también el rendimiento que ofrece en localizaciones remotas es adecuado.

En esta sección no han sido incluidas todas las experiencias de evaluación realizadas con el sistema SIETTE. Además de en la asignaturas de enseñanza oficial universitaria, este entorno ha sido utilizado también en cursos de especialización, y en las dos primeras ediciones (cursos 2003/05 y 2004/06) del *Máster de Informática aplicada a las Comunicaciones Móviles*, financiado por Fundación Vodafone junto con la Universidad de Málaga. SIETTE se utiliza en todos los módulos del máster como mecanismo de inferencia del progreso que hacen los estudiantes en su asimilación de las materias que en él se imparten.

Ciertamente, todavía es necesario incluir algunas mejoras en el sistema, como la posibilidad de modificar las respuestas a los ítems que han sido respondidos con anterioridad. También es necesario destacar que para verificar adecuadamente el sistema es requisito indispensable su utilización como lo que es: un sistema para la generación de TAI, y evaluar el impacto que este tipo de tests tienen sobre los estudiantes españoles.

La tabla 7.23 muestra un resumen de todos los tests que se han administrado a través de SIETTE como complemento docente en la enseñanza oficial universitaria hasta mayo de 2005. Los datos se han organizado por asignaturas (cada fila corresponde a una de ellas). La segunda columna indica las titulaciones en las que se imparte y la universidad a la que pertenece; la tercera representa los prototipos con los que se realizaron tests; la cuarta las convocatorias en las que se utilizó SIETTE; y por último, la quinta y la sexta el número de

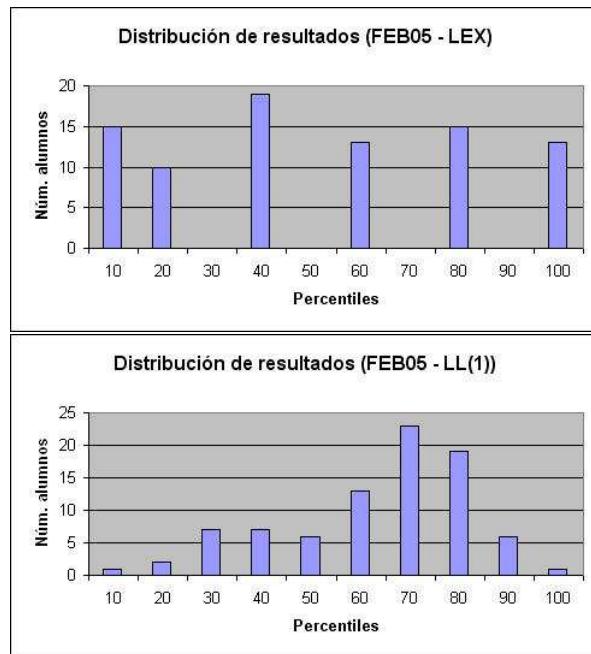


Figura 7.35: Resultados de los tests de Procesadores del Lenguaje administrados en febrero de 2005.

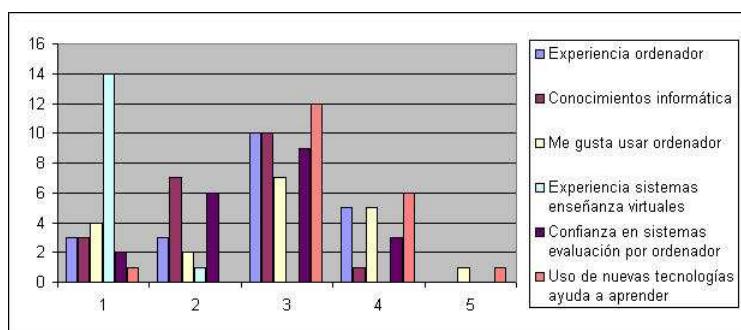


Figura 7.36: Resultados del test de botánica administrado en mayo de 2005 referentes a la experiencia de los alumnos en el uso de PC.

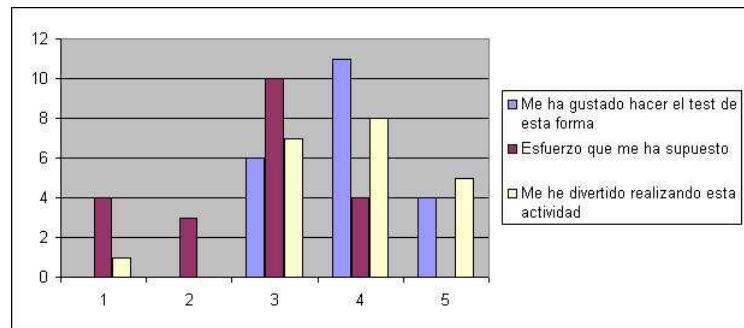


Figura 7.37: Resultados del test de botánica administrado en mayo de 2005 referentes a la valoración que los alumnos hicieron de la actividad en sí.

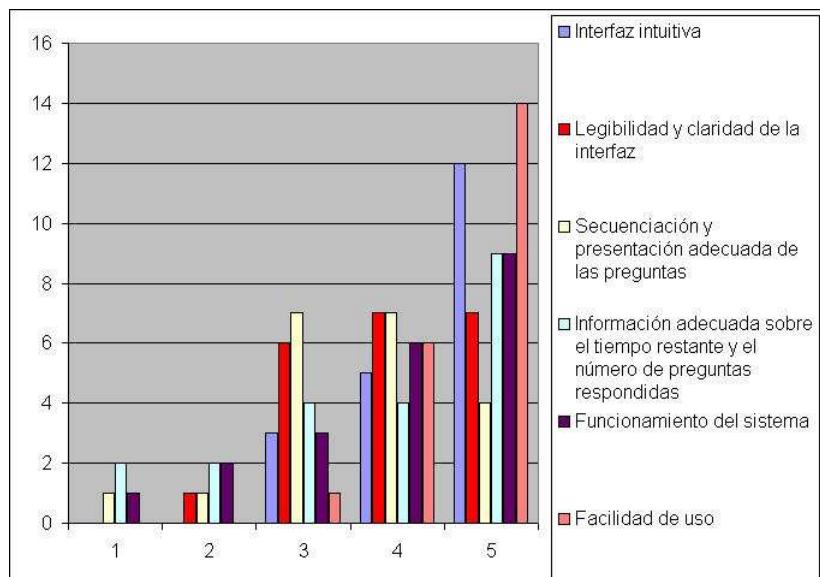


Figura 7.38: Resultados del test de botánica administrado en mayo de 2005 referentes a la valoración que los alumnos hicieron de las características de SIETTE.

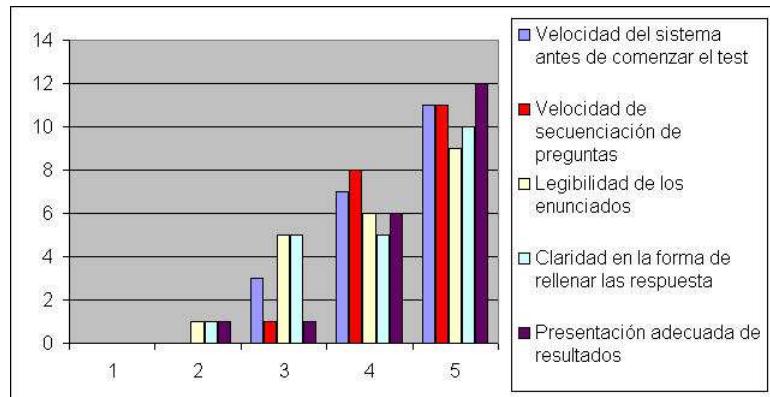


Figura 7.39: Resultados del test de botánica administrado en mayo de 2005 referentes a la valoración que los alumnos hicieron sobre las características de SIETTE.

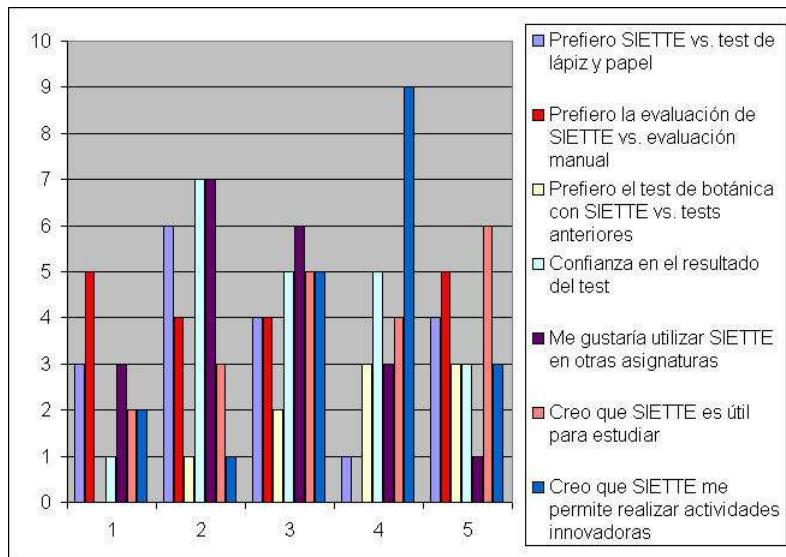


Figura 7.40: Resultados del test de botánica administrado en mayo de 2005 referentes a la valoración que los alumnos hicieron frente a las técnicas tradicionales.



Figura 7.41: Resultados del test de botánica administrado en mayo de 2005 referentes a la valoración global.

ítems y la cantidad de alumnos, respectivamente, que realizaron el test administrado en la convocatoria situada en la misma posición de la cuarta columna.

7.10. Discusión y conclusiones generales del capítulo

En este capítulo se ha llevado a cabo un estudio de los modelos de respuesta y diagnóstico cognitivo, desde la perspectiva de la evaluación sumativa. Inicialmente, se han analizado las ventajas y desventajas que supone la inclusión de los nuevos tipos de ítems que aporta el modelo de respuesta. Como resultado se ha obtenido que ésta supone una reducción significativa en cuanto al número de ítems requerido para diagnosticar el nivel conocimiento del alumno. Estas mejoras, aunque se producen independientemente del criterio de selección utilizado en la aplicación del TAI, la reducción es más significativa para el basado en la entropía (reducción del 64 %).

Se han estudiado los diferentes criterios de selección incluidos en el modelo de evaluación. De entre ellos, el que ha dado mejores resultados, en cuanto a número medio de ítems empleados en el diagnóstico, es el basado en la entropía. Asimismo, la aplicación de este criterio se traduce también en una mejora en la precisión de las estimaciones. En cuanto al número de alumnos simulados cuyo nivel ha sido correctamente diagnosticado, los porcentajes de acierto son más o menos similares en todos los mecanismos de selección.

Asimismo, se ha llevado a cabo una comparación del modelo de respuesta propuesto con el modelo clásico continuo basado en la función 3PL. Inicialmente, se ha comparado una versión dicotómica del modelo de respuesta propuesto en la tesis. Como era de esperar, en este caso, el modelo 3PL se comportaba considerablemente mejor, en cuanto a número de ítems requeridos para el diagnóstico. La diferencia es especialmente relevante cuanto menor es el valor del umbral de finalización. Por el contrario, cuando se utiliza el modelo de respuesta en su versión politómica, éste es mejor que el 3PL continuo. Además, también se ha llevado a cabo un estudio comparativo de ambos modelos en cuanto a tiempo de cómputo requerido para llevar a cabo las fases de un TAI, esto es, selección del ítem, inferencia del conocimiento del alumno y comprobación de si se satisface el criterio de finalización. Este estudio ha puesto de manifiesto, primeramente que hay un cuello de botella en la fase de selección del TAI. También se ha visto que el modelo propuesto, aún siendo politómico (a diferencia del modelo 3PL), requiere mucho menos tiempo de cómputo. Esto representa

Asignatura	Titulación/Universidad	Prototipos	Fechas	Número de ítems	Número de alumnos
Procesadores del Lenguaje	Ingeniero en Informática/UMA	1, 2, 3, 4	Feb03, Sep03, Dic03, Feb04, Sep04, Dic04, Feb05	25, 18+15, 20, 1+15, 30, 20, 5+21	85, 23, 83, 105, 39, 80, 85
Traductores, Compiladores e Intérpretes	Ingeniero Técnico en Informática de Gestión/UMA	1	Abr03	25	101
Inteligencia Artificial e Ingeniería del Conocimiento	Ingeniero en Informática/UMA	3, 4	Dic03, Feb04, Jun04, Sep04, Dic04, Ene05	20, 20, 20, 20, 20, 20	83, 44, 9, 29, 79, 54
Laboratorio de Ingeniería del Software	Ingeniero en Telecomunicaciones/UMA	3	Feb04	50	84
Laboratorio de Tecnología de Objetos	Ingeniero en Informática, Ingeniero Técnico en Informática de Gestión, Ingeniero Técnico en Informática de Sistemas/UMA	3	Jun04	[10, 10, 10]	[79, 89, 107]
Botánica	Ingeniero Técnico Forestal/UPM	4	May05	100	21

Tabla 7.23: Resumen de los tests administrados durante la evaluación formativa de SIETTE.

Leyenda: El símbolo '+' indica que en la convocatoria correspondiente se administraron dos tests (uno tras otro). Los corchetes se utilizan para representar que los resultados que incluyen son de un mismo test aplicado a grupos diferentes.

una ventaja, puesto que al haberse implementado sobre la Web, es muy importante que las diferentes etapas del TAI intenten economizar el tiempo de cómputo.

A continuación se ha estudiado si realmente los criterios adaptativos de selección de ítems incluidos en el modelo de diagnóstico son capaces por sí solos de llevar a cabo una selección balanceada de ítems cuando el test es multiconceptual. Este estudio ha sido aplicado únicamente a los criterios de selección bayesiano y basado en la entropía, puesto que los criterios restantes se han diseñado incluyendo explícitamente una fase en la que se elige el concepto cuyo diagnóstico es más impreciso. Los resultados han mostrado que efectivamente ambos criterios llevan a cabo una selección balanceada de ítems.

Posteriormente, se ha abordado el estudio del algoritmo de calibración propuesto para los ítems del modelo de respuesta. En este estudio se han llevado a cabo diversas simulaciones a partir de los resultados de someter a diversas muestras de alumnos de tamaño variable a un test (no adaptativo) evaluado heurísticamente con los ítems que se deseaba calibrar. Asimismo, por cada muestra se ha repetido el proceso de calibración para diferentes valores del parámetro de suavizado del algoritmo. Los resultados han mostrado cómo el algoritmo va reduciendo el ECM de las CCO conforme la muestra de alumnos es mayor. Además, se ha puesto de manifiesto como, en general, los mejores valores se obtienen cuando el valor del parámetro de suavizado oscila entre 0,75 y 0,85. Asimismo, los resultados han mostrado que cuando se aplican TAI, incluso con las CCO calibradas a partir de una muestra de alumnos de tamaño mínimo (20 alumnos), se obtienen un acierto en el diagnóstico del conocimiento del alumno del 90% o superior.

Se ha analizado cómo influye el tipo de función núcleo utilizada en la calibración sobre

la bondad de los resultados finales. Este estudio ha revelado que, al menos para el modelo de respuesta propuesto en esta tesis, el uso de la función gaussiana se traduce en mejores resultados en términos del error cuadrático medio cometido, y del porcentaje de individuos de la muestra cuyo nivel de conocimiento ha sido inferido correctamente.

Dentro del estudio del algoritmo de calibración, se ha analizado si el criterio de evaluación heurístico utilizado en la primera fase de calibración es relevante en los resultados finales de ésta. Con este fin, se han probado los dos criterios de evaluación heurísticos definidos en el modelo, es decir, el porcentual y el por puntos. Los resultados han puesto de manifiesto que cuando se aplica el criterio por puntos los resultados de la calibración son notablemente mejores.

La última parte del estudio del algoritmo de calibración se ha centrado en comparar la propuesta original de Ramsay con la modificación sugerida en esta tesis. Los resultados han mostrado que con la propuesta realizada en esta tesis se mejoran los resultados de la calibración en términos del ECM de las CCO calibradas, y del porcentaje de alumnos de la muestra cuyo nivel de conocimiento ha sido inferido correctamente. Este estudio también ha revelado que esa mejora final tiene la desventaja de que requiere un número de iteraciones mayor que la propuesta original.

Tanto el modelo de respuesta como el de diagnóstico han sido implementados en el sistema Web SIETTE. Por este motivo, en 2002 se realizó un estudio en el que el objetivo era evaluar un sistema Web (en este caso, SIETTE) como medio de recopilación de datos de individuos que realizan tests de forma no adaptativa para la posterior calibración de los ítems. En este estudio se comparó este medio de recolección con el convencional de papel y lápiz. El análisis mostró que los resultados de la calibración eran bastante similares, lo que permite concluir que la Web, siempre que se haga en determinadas condiciones, es un medio apto para este fin.

Finalmente, se ha realizado la evaluación formativa de SIETTE. En este estudio se han descrito los diferentes prototipos por los que ha pasado el sistema. En cada uno de ellos se han descrito las características que añadía con respecto al anterior, y los diversos experimentos que con él se han realizado. Asimismo, se han explicado también los resultados de cada experimento, intentando enfatizar los problemas surgidos. Al final de la evaluación formativa, se ha estudiado los resultados de la encuesta que se hizo a los alumnos de botánica de la Universidad Politécnica de Madrid sobre SIETTE. En general, éstos pueden considerarse satisfactorios, aunque es cierto que los alumnos siguen siendo un poco reacios al uso de las nuevas tecnologías, especialmente en la evaluación educativa.

Parte VI

CONCLUSIONES

Capítulo 8

Conclusiones

*Lo mismo es nuestra vida que una comedia;
no se atiende a si es larga,
sino a si la han representado bien.
Concluye donde quieras,
con tal de que pongas buen final.*

Lucio Anneo Séneca

Durante todos los capítulos de esta tesis se ha puesto de manifiesto la relevancia de la evaluación en el proceso de aprendizaje. Su importancia no sólo deriva de la necesidad del profesor de partir de evidencias que le permitan guiar el proceso de instrucción del estudiante, sino también de la propia necesidad de éste de cuantificar sus avances.

En la actualidad, los STI han evolucionado, principalmente, en los denominados SEAW, cuya plataforma de ubicación es Internet. En ellos, la importancia de un buen proceso de diagnóstico del conocimiento del alumno es vital, puesto que de esta información depende el proceso de aprendizaje en sí mismo. Recuérdese que a partir de las inferencias realizadas en el proceso de diagnóstico, el módulo pedagógico debe actuar en consecuencia y determinar qué paso debe seguir el estudiante en la instrucción.

En general, las soluciones que se han propuesto en la literatura de los STI se basan, o bien en procedimientos *ad hoc*, únicamente adecuados para la instrucción en una materia muy concreta, y por tanto, difícilmente aplicables a otros dominios; o bien, en el uso de tests. Estos últimos son pruebas a las que se somete a los alumnos, en las que, a través de un conjunto de ítems, se puede inferir su conocimiento de una materia. Entre las ventajas del uso de los tests, destaca su carácter genérico, puesto que pueden aplicarse a prácticamente cualquier dominio (declarativo). El principal problema que presentan los tests (convencionales), y las técnicas utilizadas para el diagnóstico en general, reside en los mecanismos empleados para realizar las inferencias, ya que los resultados que se obtienen se basan en heurísticos, que carecen de rigor teórico.

Desde el campo de la psicometría, se proponen los TAI como un intento de realizar diagnósticos basados en un fundamento teórico, la TRI. El problema principal estriba en que el proceso de construcción de un TAI es muy costoso y complejo y, en general, requiere de un conjunto de recursos que no siempre están disponibles. En el ámbito de los STI existen diversos intentos de utilizar los TAI como elementos de diagnóstico. Como se ha puesto de manifiesto, la mayoría de las propuestas de este tipo se basan en el uso de TAI que utilizan

el modelo de respuesta dicotómico 3PL. Entre las principales desventajas que presentan los TAI, desde el punto de vista del diagnóstico, una de ellas es que, en un mismo test, sólo pueden diagnosticar el conocimiento del examinando en un único concepto. Asimismo, cuando en un test se incluyen ítems que evalúan diversos conceptos, los heurísticos son el único mecanismo que consigue que la selección se haga de forma balanceada, es decir, que el número de ítems de cada concepto sea adecuado.

La solución propuesta en esta tesis para el diagnóstico en STI, se basa en la definición de un modelo de diagnóstico cognitivo basado en los TAI. Los modelos de diagnóstico cognitivo son más ambiciosos que los TAI, puesto que intentan proporcionar un conjunto de inferencias más sofisticado que la estimación numérica del conocimiento del alumno en un único concepto que se obtiene a partir de un TAI. Para dotar a este modelo de un fundamento teórico, se ha definido además un modelo de respuesta basado en la TRI. Ambos han sido implementados en un entorno Web, el sistema SIETTE, que es capaz de operar como una herramienta de diagnóstico independiente, o bien como un módulo integrable dentro de la arquitectura de un STI o un SEAW.

En este capítulo se intentan desgranar las aportaciones que supone esta tesis. Asimismo, se detallan las ventajas que ofrece y algunas de las limitaciones que presenta el modelo propuesto. Finalmente, se exponen cuáles son las líneas de investigación que se abren con este trabajo, y que a su vez suponen su propia continuación.

8.1. Aportaciones

A continuación se procederá a detallar las aportaciones que ha supuesto esta tesis. Éstas pueden clasificarse en tres grupos:

1) *Aportaciones al campo de la TRI*: Se ha definido un modelo de respuesta basado en la TRI que es, en líneas generales, politómico, discreto y no paramétrico. Entre las características más interesantes que tiene, pueden destacarse las siguientes:

- Es un modelo politómico versátil definido para un conjunto heterogéneo de tipos de ítems, y no limitado únicamente a los clásicos de opción múltiple. La mayoría de los modelos de respuesta se centran en los ítems de opción múltiple. Ciertamente, desde el punto de vista de los modelos paramétricos, la inclusión de diversos tipos de ítems dificulta tareas como la calibración. En el modelo de respuesta propuesto se han incluido no sólo ítems de opción múltiple, sino también los clásicos de verdadero/falso, y otros tipos más innovadores tales como los de respuesta múltiple con opciones dependientes o independientes, de ordenación y de relación. Todos ellos pueden combinarse libremente en un mismo test.

Para dar homogeneidad en el tratamiento de todos estos ítems, se han definido dos tipos de curvas características: las CCO y las CCR. Las primeras modelan la probabilidad que tiene un individuo de seleccionar una opción de respuesta en concreto de un ítem; mientras que las segundas, las CCR, modelan la probabilidad de elegir un determinado patrón de respuestas (una o más opciones de respuesta) del ítem. A través de estas curvas características (junto con las CCI) se ha conseguido que las diferentes fases de un TAI: calibración, selección e inferencia del conocimiento del alumno, etcétera, se lleven a cabo de forma completamente homogénea, sin distinción alguna por el tipo de ítem del que se trate.

- Es un modelo de respuesta discreto, en el que el profesor establece a priori los niveles de conocimiento en los que se evalúa al alumno, cuando construye el modelo conceptual de la asignatura. Esto implica que las curvas características se convierten en vectores de dimensión igual al número de niveles de conocimiento, y las distribuciones de conocimiento en vectores de probabilidad. Gracias a esta discretización, el nivel de conocimiento oscila entre cero (valor mínimo) y el número de niveles de conocimiento menos uno (máximo valor). Esta discretización facilita la inferencia del conocimiento del examinando a partir de la evidencia que supone la respuesta a un ítem, y reduce significativamente el tiempo requerido para calcular cuál es el ítem que debe mostrarse en cada momento, así como para determinar si el test debe finalizar. También facilita la portabilidad de las distribuciones del conocimiento del alumno entre el modelo y un STI, ya que éstas pueden ser enviadas a este último mediante pares nivel/probabilidad inferida de que lo que sabe corresponda a ese nivel.
- El modelo presentado es no paramétrico. En determinadas ocasiones la carencia de datos estadísticos para llevar a cabo la calibración de los ítems impide que las curvas características puedan ser correctamente modeladas de forma paramétrica. La razón principal es que los modelos paramétricos buscan un compromiso entre una función determinada y unos datos estadísticos que deben ajustarse a ella. Por el contrario, con el uso de modelos no paramétricos, las curvas características sólo se ajustan a los datos obtenidos a partir de sesiones de tests realizadas por alumnos con esos ítems, por lo que se puede decir que este tipo de modelos son más cercanos a la realidad.
- Se ha diseñado un proceso de calibración de ítems en el que se han aplicado técnicas estadísticas de suavizado núcleo, que requieren un número reducido de individuos para poder llevar a cabo esta tarea correctamente. Este procedimiento está inspirado en un método desarrollado con anterioridad para calibrar ítems de opción múltiple de forma no paramétrica. Se ha adaptado para ser aplicado no sólo a los ítems de opción múltiple, sino también al resto de los que proporciona el modelo de respuesta. Con respecto al algoritmo original, se han llevado a cabo diversas modificaciones que lo han mejorado.
- Este modelo de respuesta no restringe el número de opciones de respuesta que puede tener un determinado ítem. Por este motivo, se ha definido una aproximación cuasipolítómica del modelo. Esto permite, acometer adecuadamente la calibración de todas las CCO, aíñ en el caso de que la muestra poblacional no sea muy grande. Esta versión del modelo aplica *bootstrapping* a los datos, para determinar de qué respuestas se dispone de suficiente información para llevar a cabo la calibración de su CCO. De esta forma, las CCO de todas aquellas opciones que han sido seleccionadas por un porcentaje menor de alumnos de la muestra insuficiente para llevar a cabo la calibración, se tratan como una única opción de respuesta, y por tanto, sólo se les asigna una única CCO. Gracias a este mecanismo tan simple, incluso los ítems más complejos (con un número de opciones de respuesta considerablemente grandes) son abordables, sin tener que recurrir a su dicotomización. En resumen, esta aproximación cuasipolítómica, permite definir modelos basados en la información, en el sentido de que el grado de politomicidad del modelo vendrá restringido por la información estadística de la que se disponga en el momento de llevar a cabo el proceso de calibración de los ítems.

Si bien es cierto que algunas de las características del modelo están presentes por separado en otros modelos de respuesta, quizás lo novedoso de esta aportación es la combinación de ellas, y la homogeneidad en su aplicación a los distintos tipos de ítems.

- 2) *Aportaciones al diagnóstico del alumno en STI:* Utilizando como fundamento teórico subyacente el modelo de respuesta anterior, se ha definido un modelo de diagnóstico cognitivo para STI, en el que se utilizan TAI con características especiales. Este modelo, entre otras cosas, permite la evaluación simultánea de diversos conceptos en un mismo test. A modo de resumen, esta propuesta presenta las características que se detallan a continuación:
 - Dentro de la arquitectura de este modelo, se ha definido un módulo experto formado por tres componentes fundamentales para el diagnóstico. Por un lado, un modelo conceptual, que permite representar una asignatura mediante una jerarquía de conceptos, gracias a la cual se puede evaluar el conocimiento del alumno en dominios declarativos. Esta jerarquía es de granularidad variable, a determinar por el profesor que la construye. Este modelo conceptual se combina con un banco de ítems y un conjunto de especificaciones de tests, elementos indispensables para el diagnóstico del alumno.
 - El modelo de diagnóstico incluye un mecanismo de inferencia y actualización del conocimiento en múltiples conceptos (en una única sesión de evaluación), en virtud de las relaciones que se establecen entre éstos en el modelo conceptual. Mediante este mecanismo, cuando un sujeto está realizando un test, el modelo es capaz de inferir su conocimiento en los conceptos evaluados directamente, en aquéllos descendientes de éstos e incluso en otros precedentes. Gracias a esta característica, este modelo de diagnóstico podría ser adecuado como herramienta de inicialización y actualización del modelo cognitivo del alumno en un STI.
 - En este modelo se han adaptado diversos criterios de selección de ítems comúnmente utilizados en el ámbito de los TAI, tales como el método basado en la máxima información, en su versión politómica; el método bayesiano de la máxima precisión esperada, que ha sido adaptado al modelo de respuesta politómico; y el modelo basado en la dificultad. Asimismo, se ha añadido un nuevo método de selección de ítems, inspirado en el concepto de entropía de Shannon, también llamado *criterio de ganancia de información*. Este criterio, en su formulación original, era sólo apto para respuestas dicotómicas. En el modelo de diagnóstico presentado en esta tesis, este mecanismo, que se ha denominado *método basado en la entropía esperada*, se ha extendido y adaptado a las características politómicas del modelo de respuesta utilizado. Este nuevo criterio de selección es mejor que los anteriores, en cuanto a que reduce de forma significativa el número de ítems necesario para diagnosticar el conocimiento del alumno.
 - El modelo de diagnóstico cognitivo incluye un mecanismo para la realización de tests balanceados en contenido, totalmente adaptativo. Como se ha puesto de manifiesto en capítulos anteriores, el problema del balanceo en contenido en tests ha sido tratado tanto en el ámbito de la psicometría como en el de los STI. Este problema consiste en que, cuando un alumno realiza un test en el que se evalúan, de forma simultánea, múltiples conceptos, es necesario garantizar que el número de ítems de cada concepto sea suficiente. La mayoría de las propuestas para garantizar el balanceo en contenido, en ambas disciplinas, se fundamentan en el uso de heurísticos. Mediante éstos, el profesor indica (de forma manual,

según su propia experiencia) qué porcentaje de ítems de cada concepto deben incluirse en el test. Esta solución comúnmente aceptada, contradice la propia naturaleza de los TAI, ya que uno de los argumentos que se esgrime a favor de este tipo de tests frente a los convencionales es que tienen un fundamento teórico subyacente que garantiza la validez de los resultados.

La solución propuesta en esta tesis no hace uso de heurísticos, sino que se basa en el concepto de dispersión (expresado por medio de la varianza) de las distribuciones probabilísticas del conocimiento del alumno. Una distribución presenta una mayor grado de dispersión cuanto más achatada es; por el contrario será menos dispersa cuando más apuntada sea. De esta forma, y teniendo en cuenta esto, el mecanismo de balanceo en los tests multi-conceptuales se basa en dividir la selección de ítems en dos fases. En una primera, se selecciona aquel concepto cuya estimación es más dispersa, y a partir de los ítems de éste, en una segunda fase, se elige aquél que corresponda según el criterio de selección establecido en el test.

Esta división en dos fases sólo es necesaria realizarla en los criterios de selección basados en la dificultad y en la información máxima. Los métodos de elección bayesiano y basado en la entropía, tal y como se han formulado para este modelo, y por sí solos, realizan una selección balanceada en este tipo de tests, sin necesidad de añadir una fase de selección previa del concepto.

- En el modelo de diagnóstico se han definido tres criterios adaptativos para determinar cuando debe finalizar un test. Éstos pueden combinarse con otros no adaptativos, que se utilizan para garantizar la finalización de test.
 - El procedimiento de calibración del modelo de respuesta se ha extendido para poder calibrar todas las curvas características que surgen como consecuencia de la estructura del modelo conceptual del modelo de diagnóstico.
- 3) *Aportaciones desde el punto de vista de la implementación:* El modelo de diagnóstico anterior se ha implementado en una herramienta de evaluación sobre la Web: el sistema SIETTE. Entre las características de SIETTE, pueden destacarse las siguientes:
- El sistema no sólo permite la realización de tests adaptativos, sino también la administración de tests convencionales, facilitando de esta forma el soporte necesario para poder llevar a cabo la calibración de los ítems, sin necesidad de recurrir a otras herramientas. Ha sido utilizado en la práctica en la E.T.S.I. Informática, en la E.T.S.I. Telecommunicaciones y en el Máster Universitario de Informática aplicada a las Comunicaciones Móviles de la Universidad de Málaga; y en la E.U. de Ingeniería Técnica Forestal de la Universidad Politécnica de Madrid.
 - SIETTE permite incluir ítems que requieren una interacción mayor que los habituales, en los que el examinando tan sólo debe seleccionar una o más opciones de respuesta. Mediante los ítems autocorregidos (o *siettlets*) de SIETTE, se pueden incluir prácticamente cualquier tipo de ejercicio que pueda evaluarse como uno de los tipos de ítems del modelo de respuesta. La única restricción sobre estos ejercicios más sofisticados es que deben implementarse mediante lenguajes de programación que puedan ser embebidos en una página HTML. En esta línea, se ha construido una biblioteca para la construcción de ejercicios que permite, a aquellos profesores sin conocimiento de programación, combinar en sus tests ítems tradicionales con otros más sofisticados.

- Asimismo, se han definido los denominados ítems externos, que son aquéllos cuyo contenido no reside en la base de conocimiento de SIETTE. Para este tipo de ítems, SIETTE sólo almacena sus propiedades psicométricas, junto con una URL a través de la cual se invoca al ítem en el momento de la presentación.
- Incluye un mecanismo simple de generación de ítems isomorfos, lo que facilita al profesor la tarea de construir nuevos ítems. Éstos se basan en la inclusión de sentencias de generación pseudoaleatoria de números en el enunciado y en las opciones de respuesta. Los ítems generativos podrán ser a su vez ítems de cualquiera de los tipos que incluye el modelo de respuesta.
- El sistema incluye un cierto carácter tutorial al permitir añadir en cada ítem, refuerzos junto con los diferentes patrones de respuesta, y un conjunto de pistas cuyo objetivo es ayudar al alumno a resolverlo. Los refuerzos y las ayudas permiten construir los denominados tests de autoevaluación, a través de los cuales los alumnos se ven envueltos en un proceso de aprendizaje socrático.
- Se ha puesto de manifiesto que la Web es un medio para la recolección de evidencias para la calibración de ítems, tan válido con los tests de papel y lápiz. Ciertamente, la única restricción que se debe imponer es que los individuos que participen en el test utilizado para la recolección de datos estén bajo la supervisión de los profesores en el momento de su realización.
- Finalmente, es un sistema que puede funcionar de forma autónoma para autoevaluación o evaluación, o como módulo de diagnóstico integrable dentro de STI, gracias a un conjunto de protocolos que permiten, de forma simple, que ambos sistemas interactúen y se intercambien información. Estos protocolos permiten asimismo, la integración a distintos grados de acoplamiento, permitiendo incluso a SIETTE asumir el rol de un servicio Web para el diagnóstico del conocimiento del alumno. Como consecuencia, SIETTE ha sido integrado en sistemas como TAPLI y TRIVIETTE, y en la actualidad está siendo integrado en otros como LeActiveMath y MEDEA.

8.2. Limitaciones

Toda propuesta tiene sus limitaciones. Si bien es cierto el modelo de evaluación presentado en esta tesis contribuye en la resolución de algunos de los problemas existentes en el diagnóstico del conocimiento del alumno, y en el uso de TAI para la construcción de modelos de diagnóstico cognitivos, éste presenta las limitaciones siguientes:

- Permite administrar los denominados *tests con multidimensionalidad entre ítems* (en inglés, *between-item multidimensional tests*) (Wang y Chen, 2004). Este grado de multidimensionalidad se produce cuando en un mismo test coexisten ítems unidimensionales que evalúan diferentes conceptos. Por el contrario, tal y como ha sido formulado en esta tesis, este modelo no permite actualmente los denominados *tests con multidimensionalidad intra-ítems* (en inglés, *within-item con multidimensionalidad tests*). Este tipo de tests son los que se basan en modelos de respuesta multidimensionales, en los que un mismo ítem proporciona evidencias sobre el conocimiento del alumno en más de un concepto, siendo estos conceptos independientes entre sí.
- Teorías como el constructivismo (Piaget, 1952; Vygotskii, 1978; Minsky, 1986) postulan que el progreso de un estudiante no sólo se manifiesta a través de un resultado

satisfactorio en un test sobre los conceptos estudiados. También otros logros importantes del alumno, tales como una mejora en la colaboración, la auto-regulación, el auto-control y otras habilidades importantes, no pueden juzgarse únicamente a partir de su éxito en la realización de un ejercicio. El modelo de diagnóstico propuesto es cognitivo, y por tanto, únicamente es capaz de medir el conocimiento del alumno en un conjunto de conceptos.

- Siguiendo en la misma línea, la evaluación mediante tests es únicamente apta para dominios estructurados en una jerarquía de conceptos, y por lo tanto no es para evaluar dominios procedimentales. También es cierto que, en el ámbito del diagnóstico procedural, no existen herramientas genéricas que estén además basadas en un fundamento teórico que justifique la validez de sus inferencias, a diferencia del modelo de diagnóstico presentado en esta tesis.
- El modelo conceptual incluido en esta propuesta sólo contempla relaciones de agrupación entre los conceptos. No incluye por tanto, otro tipo de relaciones, como por ejemplo, la de prerrequisito. En la actualidad existen propuestas de modelos basados en la TRI, como la realizada en (Pérez de la Cruz et al., 2005), en la sí se tienen en cuenta este tipo de relaciones.
- El sistema SIETTE permite la posibilidad de incluir ítems generativos en un test. Desde el punto de vista psicométrico, en SIETTE, éstos se tratan de forma análoga al resto de ítems, sin hacer ningún tipo de distinción. Además, en el proceso de calibración no se han desarrollado mecanismos específicos para tratar este tipo de ítems. En este sentido, bien es cierto que trabajos como el realizado por Béjar et al. (2003) demuestran que durante la calibración, no es necesario hacer distinción en el caso de que los ítems sean generativos.

8.3. Líneas de investigación abiertas

Este trabajo no supone el final de una línea de investigación. Por el contrario, no es más que el principio de un largo camino frente al cual se abren nuevas líneas de investigación, algunas de las cuales se enumeran a continuación:

- Desarrollo de una versión del modelo de respuesta con multidimensionalidad intra ítem. Ciertamente, en la actualidad, SIETTE tiene implementada la funcionalidad básica necesaria para mantener y utilizar curvas características y distribuciones de conocimiento del alumno multidimensionales. Asimismo, posee el soporte necesario para considerar la multidimensionalidad en las distintas fases del algoritmo de un TAI, esto es, selección del ítem, actualización e inferencia del conocimiento del alumno y determinación de si el test debe finalizar. A pesar de estar implementadas, estas funcionalidades no han sido probadas para determinar su factibilidad y validez. Igualmente, tampoco se ha desarrollado una versión del algoritmo de calibración que permita inferir las curvas características de los ítems multidimensionales.
- La inclusión de ítems con formatos diferentes es un aspecto que debería ser estudiado con detalle. Sería interesante analizar si existe alguna relación entre el tipo de formato de los ítems del test y el nivel de conocimiento del alumno (¿Mejoran los resultados de una muestra de individuos si los ítems del test tiene un formato más atractivo y sofisticado?). En caso de existir alguna relación entre formato y nivel de conocimiento,

quizás podría añadirse un nuevo nivel a la selección adaptativa de ítems (durante el test), en la que se determinara en qué formato debería mostrarse éstos, dado el conocimiento estimado actual del examinando (Eskenasi et al., 1992).

- El tiempo de respuesta a un ítem es otro factor interesante de estudiar. En la actualidad existen diversos modelos basados en la TRI que tienen en cuenta el tiempo de respuesta como parte integrante de la propia respuesta. Podría ser interesante extender el modelo de respuesta propuesto en esta tesis para contemplar esta información adicional.
- Utilización de modelos basados en la TRI para la construcción de un planificador de instrucción para STI que determine cuál es la acción que debe llevar a cabo el estudiante para mejorar la calidad de su proceso de aprendizaje. Si los criterios de selección de ítems permiten seleccionar qué ítem es más adecuado para el alumno, en función de su conocimiento actual, por qué no utilizar este mismo razonamiento para determinar, según el conocimiento actual del sujeto, cuál es la estrategia tutorial más adecuada.
- El algoritmo de calibración utilizado en esta tesis, permite calibrar ítems que no lo han sido con anterioridad. Otra línea interesante de investigación sería el desarrollo de una nueva versión del algoritmo que permitiese la calibración de los ítems en línea. Aunque obviamente la versión actual permite volver a calibrar los ítems con nuevas evidencias; este proceso, tal cual, requiere recopilar todas las evidencias utilizadas para la calibración inicial, unirlas con las nuevas evidencias, y repetir el proceso partiendo desde cero. La idea que subyace en la calibración en línea es, a partir de nuevas evidencias, y a partir de las curvas características calibradas, incorporar dinámicamente las nuevas evidencias para mejorar la estimación de las curvas características (Conejo et al., 2000).

Siguiendo con la calibración, sería interesante incluir en el algoritmo de calibración propuesto, alguna técnica especial de atenuación para la calibración de los ítems generativos, como la propuesta en (Mislevy et al., 1994), que tuviese en cuenta las características de este tipo de ítems.

- La inclusión de ayudas y refuerzos dota a SIETTE de un cierto carácter tutorial. Por cada ítem, es posible incluir un conjunto de ayudas. Una extensión del modelo de diagnóstico cognitivo sería la inclusión de un mecanismo de selección adaptativa de la ayuda más adecuada al nivel de conocimiento del examinando que la solicita. De igual forma, en esta extensión del modelo, el uso de una ayuda debería tenerse en cuenta a la hora de actualizar la estimación del conocimiento del individuo. En esta línea ya se han realizado algunos avances (Conejo et al., 2003, 2005).

En relación con esta línea de trabajo, y como se mencionó en el capítulo 6, la inclusión de refuerzos podría violar una de las hipótesis de la TRI, la invarianza (véase sección 2.8). Este problema ha sido tratado por diversos investigadores (Embretson, 1991, 1993; Klauer y Sydow, 2000; Conati et al., 2002) que han contemplado la posibilidad de que el alumno aprenda durante el test. Una línea de trabajo abierta es el desarrollo de un modelo que permita que los examinandos aprendan durante el test. Esto convertiría a una TAI no sólo en una herramienta de evaluación, sino también en una herramienta de instrucción socrática.

- Otra línea interesante de investigación, a la vez que compleja, es el uso de modelos basados en la TRI para el diagnóstico procedimental. Las propuestas que existen en

la actualidad (Martin y Mitrovic, 2002; Ferrero, 2004) para este tipo de diagnóstico, se basan, principalmente, en el uso de lenguajes lenguajes declarativos como CLIPS, o en redes bayesianas, pero no están sustentadas por una base teórica como la TRI, que garantice la fiabilidad y validez de los diagnósticos realizados.

- Por último, el modelo propuesto podría extenderse con el objetivo de ser emplearse dentro de un entorno de evaluación colaborativo en el que diversos examinandos puedan trabajar en grupo para resolver un mismo test, especialmente en tests de auto-evaluación. Esta herramienta permitiría estudiar si el trabajo en grupo proporciona ventajas frente a la formación individualizada.

APÉNDICES

Apéndice A

Lista de abreviaturas

1PL	Función logística de un parámetro
2PL	Función logística de dos parámetros
3PL	Función logística de tres parámetros
ALICE	Adaptive Link Insertion in Concept-based Educational System
CBAT-2	Content-Balanced Adaptive Testing
CCF	Curva Característica de Funcionamiento
CCI	Curva Característica del Ítem
CCO	Curva Característica de Opción
CCR	Curva Característica de Respuesta
CLARISSE	Clusters and Rules Issued
CRC	Curva de Respuesta Categórica
DCG	Dynamic Course Generation
EAP	Esperanza a Posteriori
ECM	Error Cuadrático Medio

ELM-ART	Episodic Learner Model Adaptive Remote Tutor
EM	Esperanza/Maximización
IA	Inteligencia Artificial
INSPIRE	Intelligent System for Personalized Instruction in a Remote Environment
MAP	Máximo a Posteriori
OLAE	On-Line/Off-line Assessment for Expertise
PASS	Personalized Assessment Module
PC	Computador Personal (del inglés, Personal Computer)
POLA	Probabilistic On-Line Assessment
SEAO	Sistema de Enseñanza Asistida por Ordenador
SEAW	Sistema Educativo Adaptativo para la Web
SIETTE	Sistema Inteligente de Evaluación mediante Tests para Teleeducación
STI	Sistema Tutor (o Instructor) Inteligente
TAAI	Test Auto-Adaptativo Informatizado
TANGOW	Task-based Adaptive Learner Guidance on the Web
TAI	Test Adaptativo Informatizado
TCT	Teoría Clásica de los Tests
TRI	Teoría de Respuesta al Ítem
TRIADS	Tripartite Interactive Assessment Development System
URL	Universal Resource Location

Bibliografía

- Abrahamowicz, M. y Ramsay, J. O. (1992). Multicategorical spline model for item response theory. *Psychometrika*, 57(1), 5–27.
- Aimeur, E., Blanchard, E., Brassard, G. y Gambs, S. (2001). Quanti: A multidisciplinar knowledge-based system for quantum information processing. En *Proceedings of international conference on computer aided learning in engineering education. calie'01* (pp. 51–57).
- Aimeur, E., Brassard, G., Dufort, H. y Gambs, S. (2002). Clarisse: A machine learning tool to initialize student models. En S. A. Cerri, G. Gouardères y F. Paraguacu (Eds.), *Proceedings of the 6th international conference on intelligent tutoring systems (its 2002). lecture notes in computer science* (pp. 718–728). New York: Springer Verlag.
- Anderson, J. R. (1988). The expert module. En *Foundations of intelligent tutoring systems* (pp. 21–54). Lawrence Erlbaum Associates, Inc.
- Anderson, J. R., Conrad, F. G. y Corbett, A. T. (1989). Skill acquisition and the lisp tutor. *Cognitive Science*, 13, 467–505.
- Anderson, J. R., Corbett, A. T., Koedinger, K. y Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2), 167–207.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Angelo, T. A. y Cross, K. P. (1993). *Classroom assessment techniques: A handbook for college teachers* (2nd ed.). San Francisco: Jossey-Bass.
- Arroyo, I., Beck, J. E., Schultz, K. y Woolf, B. P. (1999). Piagetian psychology in intelligent tutoring systems. En *Proceedings of the 8th world conference of artificial intelligence and education aied'99* (pp. 600–602). Amsterdam: IOS Press.
- Arroyo, I., Beck, J. E., Woolf, B. P., Beal, C. R. y Schultz, K. (2000). Macroadapting animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism. En *Proceedings of the 5th world conference of intelligent tutoring systems. its'00* (pp. 604–614). Springer-Verlag.
- Arroyo, I., Conejo, R., Guzmán, E. y Woolf, B. P. (2001). An adaptive web-based component for cognitive ability estimation. En *Proceedings of the 9th world conference of artificial intelligence and education aied'01*. Amsterdam: IOS Press.
- Assessment Systems Corporation. (2004a, April). *C-quest*. <http://www.assessment.com>.

- Assessment Systems Corporation. (2004b, April). *Fasttest*. <http://www.assessment.com>.
- Assessment Systems Corporation. (2004c, April). *Microcat*. <http://www.assessment.com>.
- Baker, F. B. (2001). *The basics of item response theory* (2nd ed.).
- Barbero, M. I. (1996). Bancos de ítems. En J. Muñiz (Ed.), *Psicometría* (pp. 139–170). Madrid: Universitas.
- Barbero, M. I. (1999). Gestión informatizada de bancos de ítems. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests informatizados: Fundamentos y aplicaciones* (pp. 63–83). Pirámide.
- Bardes, B. y Denton, J. (2001). Using the grading process for departmental and program assessment. En *Paper presented at the american association for higher education conference*.
- Barr, A. y Feigenbaum, E. A. (1982). *The handbook of artificial intelligence*. Pitman books limited.
- Barros, B. (1999). *Aprendizaje colaborativo en enseñanza a distancia: Entorno genérico para configurar, realizar y analizar actividades en grupo*. Tesis doctoral no publicada, Universidad Politécnica de Madrid, Madrid.
- Belmonte, M. V., Berbel, J. y Conejo, R. (1996a). Tea: An agrarian economy instructor system. En A. D. de Ilarraza Sánchez y I. F. de Castro (Eds.), *Lecture notes in computer science 1108. proceedings of the 3rd international conference on computer aided learning and instruction in science and engineering. calisce 1996* (pp. 322–330). New York: Springer Verlag.
- Belmonte, M. V., Berbel, J. y Conejo, R. (1996b). Tea: An agrarian instructor system. *European Journal of Engineering Education*, 22(4), 389–399.
- Belmonte, M. V., Guzmán, E., Mandow, L., Millán, E. y Pérez de la Cruz, J. L. (2002). Automatic generation of problems in web-based tutors. En L. C. Jain, R. J. Howlett, N. S. Ichalkaranje y G. Tonfoni (Eds.), *Virtual environments for teaching and learning* (pp. 237–281). London: World Scientific.
- Binet, A. y Simon, T. A. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *l'Année Psychologie*, 11, 191–336.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's mental ability. En *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Béjar, I., Lawless, R. R., Morley, M. E., Bennett, R. E. y Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment*, 2(3).
- Black, P. y William, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–74.
- Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals. handbook i. cognitive domain*. Longmans.

- Bloom, B. S., Hastings, J. T. y Madaus, F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51.
- Bock, R. D. (1997). The nominal categories model. En W. J. van der Linden y R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 33–49). New York: Springer Verlag.
- Bock, R. D. y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 45(4), 443–459.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.
- Boyle, A., Hutchison, D., O'Hare, D. y Patterson, A. (2002). Item selection and application in higher education. En *Proceedings of the 6th international computer assessment conference (caa)*. loughborough university (pp. 269–284).
- Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, 47, 139–159.
- Brusilovsky, P. (1998). Adaptive educational systems on the world-wide-web: A review of available technologies. En *Workshop in intelligent tutoring systems on the web. 4th international conference on intelligent tutoring system. its'98. san antonio (texas)*.
- Brusilovsky, P. (2001). Adaptive hypermedia. *User Modeling and User-adapted Interaction*, 11(1-2), 87–110.
- Brusilovsky, P. y Miller, P. (1999). Web-based testing for distance education. En *Proceedings of webnet'99, world conference of the www and internet* (pp. 149–154). AACE.
- Brusilovsky, P., Schwarz, E. y Weber, G. (1996a). Elm-art: An intelligent tutoring system on world wide web. En *Lecture notes in computer science* (Vol. 1086). Springer-Verlag.
- Brusilovsky, P., Schwarz, E. y Weber, G. (1996b). A tool for developing adaptive electronic textbooks on www. En *Proceedings of webnet'96 - world conference of the web society. san francisco (california)* (pp. 64–69).
- Brusilovsky, P. y Weber, G. (1996). Collaborative example selection in an intelligent example-based programming environment. En *Proceedings of international conference on learning sciences. ics'96* (pp. 357–362).
- Bull, S. y Pain, H. (1995). Did i say what i think i said, and do you agree with me? inspecting and questioning the student model. En *Proceedings of 7th conference on artificial intelligence in education. aied'95* (pp. 501–508).
- Burket, G. R. (1988). *Itemsys [computer program]*.
- Burket, G. R. (1991). *Pardux [computer program]*.
- Burns, H. L. y Capps, C. G. (1988). Foundations of intelligent tutoring systems: An introduction. En M. C. Polson y J. J. Richardson (Eds.), *Foundations of intelligent tutoring systems*. Lawrence Erlbaum Associates, Inc.

- Carmona, C., Bueno, D., Guzmán, E. y Conejo, R. (2002). Sigue: Making web courses adaptive. En P. D. Bra, P. Brusilovsky y R. Conejo (Eds.), *Proceedings of the 2nd international conference on adaptive hypermedia and adaptive web-based systems. ah 2002. lecture notes in computer science* (pp. 376–379). New York: Springer Verlag.
- Carmona, C., Guzmán, E., Bueno, D. y Conejo, R. (2002). An adaptive web-based tutorial of agrarian economy. En *Proceedings of the workshop on adaptive systems for web-based education, held in conjunction with ah 2002* (pp. 57–67).
- Carpenter, G. A. y Grossberg, S. (1988). The art of adaptive pattern recognition by a self-organizing neural network. *Computer*, 21(3), 77–88.
- Carro, R. M. (2001). *Un mecanismo basado en tareas y reglas para la creación de sistemas hipermedia adaptativos aplicados a la educación a través de internet*. Tesis doctoral no publicada, Universidad Autónoma de Madrid.
- Carro, R. M., Pulido, E. y Rodríguez, P. (2001). Tangow: A model for internet based learning. *International Journal of Continuing Engineering Education and Life-Long Learning, IJCEELL*, 11(Special Issue on "Internet based learning and the future of education").
- Chipman, S., Nichols, P. D. y Brennan, R. L. (1995). Introduction. En P. Nichols, S. Chipman y R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 1–18). Hillsdale, NJ: Lawrence Erlbaum.
- Chua Abdullah, S. (2003). *Student modelling by adaptive testing - a knowledge-based approach*. Tesis doctoral no publicada, University of Kent, Canterbury.
- Coll, C. (1997). *Psicología y currículum* (Sexta ed.). Ediciones Paidos Ibérica S.A.
- Collins, J. A. (1996). *Adaptive testing with granularity*. Tesis de maestría no publicada, University of Saskatchewan.
- Collins, J. A., Greer, J. E. y Huang, S. X. (1996). Adaptive assessment using granularity hierarchies and bayesian nets. En C. Frasson, G. Gauthier y A. Lesgold (Eds.), *Proceedings of the 3rd international conference on intelligent tutoring systems. its 1996. lecture notes in computer science* (pp. 569–577). New York: Springer Verlag.
- Conati, C., Gertner, A. y VanLehn, K. (2002). Using bayesian networks to manage uncertainty in student modeling. *Journal of Artificial Intelligence in Education*, 12(4), 371–417.
- Conati, C., Gertner, A., VanLehn, K. y Druzdzel, M. (1997). On-line student modelling for coached problem solving using bayesian networks. En *Proceedings of the 6th international conference on user modelling. um'97* (pp. 231–242). New York: Springer Verlag.
- Conati, C. y VanLehn, K. (1996). Pola: A student modeling framework for probabilistic on-line assessment of problem solving performance. En *Proceedings of the 5th international conference on user modeling. um'96* (pp. 75–82). User Modeling Inc.
- Conejo, R. y Guzmán, E. (2002). Siette: Un sistema inteligente de evaluación mediante tests basado en la teoría de respuesta al Ítem. *Revista de Metodología de las Ciencias del Comportamiento*, 133–137.

- Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez de la Cruz, J. L. y Ríos, A. (2004). Siette: a web-based tool for adaptive testing. *Journal of Artificial Intelligence in Education*, 14, 29–61.
- Conejo, R., Guzmán, E. y Pérez de la Cruz, J. L. (2003). Towards a computational theory of learning in an adaptive testing environment. En U. Hoppe, F. Verdejo y J. Kay (Eds.), *Artificial intelligence in education: Shaping the future of learning through intelligent technologies. (aied 2003)* (pp. 398–400). Amsterdam: IOS Press.
- Conejo, R., Guzmán, E., Pérez de la Cruz, J. L. y Millán, E. (2005). Introducing adaptive assistance in adaptive testing. En C. K. Looi, G. McCalla, B. Bredeweg y J. Breuker (Eds.), *Artificial intelligence in education: Supporting learning through intelligent and socially informed technology. (aied 2005)* (pp. 777–779). Amsterdam: IOS Press.
- Conejo, R., Millán, E., Pérez de la Cruz, J. L. y Trella, M. (2000). An empirical approach to on-line learning in siette. En G. Gauthier, C. Frasson y K. VanLehn (Eds.), *Lecture notes in computer science. proceedings of 3rd international conference on intelligent tutoring systems. its 2000* (pp. 604–614). Springer Verlag.
- Cooley, W. y Lohnes, P. R. (1976). *Evaluation research in education*. New York: John Wiley.
- Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record*, 64, 672–683.
- CTB McGraw-Hill. (2004, April). *Terranova: The second edition*.
- De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement*, 16, 327–343.
- De la Garza Vizcaya, E. L. (2004). La evaluación educativa. *Revista Mexicana de Investigación Educativa*, 9(23), 807–816.
- de Landsheere, G. (2004). *La investigación educativa en el mundo* (IV ed.). Fondo de Cultura Económica.
- Deepwell, F. (2002). Towards capturing complexity: an interactive framework for institutional evaluation. *Educational Technology & Society*, 5(3), 83–90.
- DiBello, L. V., Stout, W. F. y Roussos, L. A. (1995). Unified cognitive/psychometric diagnosis assessment likelihood-based classification techniques. En P. Nichols, S. Chipman y R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–390). Hillsdale, NJ: Lawrence Erlbaum.
- Dimitrova, V., Self, J. y Brna, P. (2001). Applying interactive open learner models to learning technical terminology. En M. Bauer, P. J. Gmytrasiewicz y J. Vassileva (Eds.), *Proceedings of the 8th international conference on student modeling (um 2001). lecture notes in computer science*. New York: Springer Verlag.
- Dodd, B. G., De Ayala, R. J. y Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*, 19(1), 5–22.
- Douglas, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika*, 62, 7–28.

- Douglas, J. (1999). *Asymptotic identifiability of nonparametric item response models* (Technical Report No. 142). University of Wisconsin. Department of Biostatistics and Medical Informatics.
- Douglas, J. y Cohen, A. (2001). Nonparametric item response function estimation for assessing parametric model fit. *Applied Psychological Measurement*, 25(3), 234–243.
- Drake Kryterion. (2004, April). *Webassessor*. <http://www.webassessor.com>.
- Drasgow, F. (1995). Introduction to the polytomous irt. special issue. *Applied Psychological Measurement*, 19(1), 1–3.
- Drasgow, F., Levine, M. V., Tsien, S., Williams, B. y Mead, A. D. (1992). Fitting polytomous item response theory models to multiple-choice tests. *Applied Psychological Measurement*, 19, 143–165.
- Embretson, S. E. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, 11, 175–193.
- Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56(3), 495–515.
- Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. En N. Frederiksen, R. J. Mislevy y I. I. Béjar (Eds.), *Test theory for a new generation of tests* (pp. 125–150). Lawrence Erlbaum Associates.
- Embretson, S. E. y Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- EQL International Ltd. (2003, April). *I-assess*. <http://www.iassess.com>.
- Escudero, T. (2003). Desde los tests hasta la investigación evaluativa actual. un siglo, el xx, de intenso desarrollo de la evaluación en educación. *Revista Electrónica de Investigación y Evaluación Educativa (RELIEVE)*. http://www.uv.es/RELIEVE/v9n1/RELIEVEv9n1_1.htm, 9(1).
- Eskenası, A., Vladimirova, T. y Vassileva, J. (1992). Incorporating student models in adaptive testing systems. *ETTI*, 30(2), 135–142.
- Eubank, R. (1988). *Spline smoothing and nonparametric regression*. New York: Dekker.
- Falmagne, J. C., Doignon, J. P., Koppen, M., Villano, N. y Johannessen, L. (1990). Introduction knowledge spaces - how to build, test and search them. *Psychological Review*, 97(2), 201–224.
- Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3), 357-381.
- Ferrando, P. J. (2004). Kernel smoothing estimation of item characteristic functions of continuous personality items: An empirical comparison with the linear and the continuous-response models. *Applied Psychological Measurement*, 28(2), 95–109.
- Ferrero, B. (2004). *Detective: un entorno genérico e integrable para diagnóstico de actividades de aprendizaje. fundamentos, diseño y evaluación*. Tesis doctoral no publicada, Facultad de Informática. Universidad del País Vasco.

- Flaugher, R. (1990). Item pools. En H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 51–65). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.
- Fostnot, C. T. (1996). *Constructivism: Theory, perspectives, and practice*. College Press.
- Friedman-Hill, E. (1997). *Jess, the java expert system shell* (Informe Técnico No. SAND98-8206). Sandia National Laboratories.
- García, E., Gil, J. y Rodríguez, G. (1998). La evaluación de tests adaptativos informatizados. *Revista Electrónica de Investigación y Evaluación Educativa (RELIEVE)*. http://www.uv.es/RELIEVE/v4n2/RELIEVEv4n2_6.htm, 4(2).
- Gaviria, J. (2002). *Breve introducción a la psicometría. principales teorías*. http://personal.telefonica.terra.es/web/drgaviria/articulos_archivos/introirt.pdf.
- Glas, C. A. W. (2000). Item calibration and parameter drift. En W. J. van der Linden y C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 138–199). Kluwer Academic Publisher.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: some questions. *American Psychologists*, 18, 519–521.
- Gálvez, J., Guzmán, E. y Conejo, R. (2003). *Una herramienta para el intercambio de especificaciones de tests entre el estándar ims y siette*. Proyecto de fin de carrera, E.T.S.Ingeniería Informática. Dpto. Lenguajes y Ciencias de la Computación. Universidad de Málaga.
- Gonçalves, J. P. (2004). *A integração de testes adaptativos informatizados e ambientes computacionais de tarefas para o aprendizado do inglês instrumental*. Tesis doctoral no publicada, USP, São Carlos (Brasil).
- Gonçalves, J. P., Aluisio, S. M., de Oliveira, L. H. M. y Oliveira, O. N. (2004). A learning environment for english for academic purposes based on adaptive tests and task-based systems. En J. C. Lester, R. M. Vicari y F. Paraguaçu (Eds.), *Proceedings of the 7th international conference on intelligent tutoring systems (its 2004). lecture notes in computer science* (pp. 1–11). New York: Springer Verlag.
- Gonvindarajulu, Z. (1999). *Elements of sampling theory and methods*. Prentice-Hall.
- Gouli, E., Kornilakis, H., Papanikolaou, K. A. y Grigoriadou, M. (2001). Adaptive assessment improving interaction in an educational hypermedia system. En *Human computers interaction. panhellenic conference with international participation* (pp. 217–222).
- Gouli, E., Papanikolaou, K. A. y Grigoriadou, M. (2002). Personalizing assessment in adaptive educational hypermedia systems. En *Lecture notes in computer science. proceedings of the 2nd international conference on adaptive hypermedia and adaptive web-based systems. ah2002* (Vol. 2347). New York: Springer Verlag.
- Greer, J. E. y McCalla, G. (1994). Granularity-based reasoning and belief revision in student models. En J. E. Greer y G. McCalla (Eds.), *Student modelling: The key to individualized knowledge-based instruction* (Vol. 125, pp. 39–62). New York: Springer Verlag.

- Grigoriadou, M., Kornilakis, H., Papanikolaou, K. A. y Magoulas, G. D. (2002). Fuzzy inference for student diagnosis in adaptive educational hypermedia. En *Methods and applications of artificial intelligence. lecture notes in artificial intelligence* (Vol. 2308). New York: Springer Verlag.
- Gulliksen, H. (1961). Measurement of learning and mental abilities. *Psychometrika*, 26, 93–107.
- Gutiérrez, J., Pérez, T. A., López Cuadrado, J., Arrubarrena, R. M. y Vadillo, J. A. (2002). Evaluación en sistemas hipermedia adaptativos. *Revista de Metodología de las Ciencias del Comportamiento, Volumen especial*, 279–283.
- Guzmán, E. y Conejo, R. (2002a). Simultaneous evaluation of multiple topics in siette. En S. Cerri, G. Gouardères y F. Paraguacú (Eds.), *Proceedings of the 6th international conference on intelligent tutoring systems (its 2002). lecture notes in computer science* (pp. 739–748). New York: Springer Verlag.
- Guzmán, E. y Conejo, R. (2002b). An adaptive assessment tool integrable into internet-based learning systems. En A. M. Vilas, J. A. M. González y I. S. de Zaldívar (Eds.), *Educational technology: International conference on tic's in education* (pp. 139–143). Badajoz: Sociedad de la Información.
- Guzmán, E. y Conejo, R. (2003a). *Using siette as a knowledge assessment service: The communication protocol* (Informe Técnico No. 2003-5). http://www.lcc.uma.es/SIETTE/docs/protocol_v1.pdf: Dpto. Lenguajes y Ciencias de la Computación. Universidad de Málaga.
- Guzmán, E. y Conejo, R. (2003b). *Xml schema specification of the interchange file of siette* (Informe Técnico No. 2003-4). <http://www.lcc.uma.es/SIETTE/docs/especificacionXML.pdf>: Dpto. Lenguajes y Ciencias de la Computación. Universidad de Málaga.
- Guzmán, E. y Conejo, R. (2004a). A model for student knowledge diagnosis through adaptive testing. En J. C. Lester, R. M. Vicari y F. Paraguacú (Eds.), *Proceedings of the 7th international conference on intelligent tutoring systems (its 2004). lecture notes in computer science* (pp. 12–21). New York: Springer Verlag.
- Guzmán, E. y Conejo, R. (2004b). A brief introduction to the new architecture of siette. En P. D. Bra y W. Nejdl (Eds.), *Proceedings of the iiith international conference on adaptive hypermedia and adaptive web-based systems(ah 2004). lecture notes in computer science* (pp. 405–408). New York: Springer Verlag.
- Guzmán, E. y Conejo, R. (2004c). A library of templates for exercise construction in an adaptive assessment system. *Technology, Instruction, Cognition and Learning (TICL)*, 2(1-2), 21–43.
- Guzmán, E. y Conejo, R. (2005). Towards efficient item calibration in adaptive testing. En L. Ardissono, P. Brna y A. Mitrovic (Eds.), *Proceedings of the 10th international conference on user modeling (um 2005). lecture notes in artificial intelligence* (pp. 414–418). New York: Springer Verlag.
- Guzmán, E. y Conejo, R. (en prensa). Self-assessment in a feasible, adaptive web-based testing system. *IEEE Transactions on Education*.

- Guzmán, E., Conejo, R. y García-Hervás, E. (2005). An authoring environment for adaptive testing. *Journal of Educational Technology & Society. Special Issue on Authoring of Adaptive Hypermedia*, 8(3), 66–76.
- Guzmán, E., Conejo, R., Hontangas, P. M., Olea, J. y Ponsoda, V. (2002). A comparative study of irt and classical item parameter estimates web-based and conventional test administration. En *International test commision's conference on computer-based testing and the internet*. Winchester (United Kingdom).
- Habing, B. (2001). Nonparametric regression and the parametric bootstrap for local dependence assessment. *Applied Psychological Measurement*, 25(3), 221–233.
- Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the dose is subject to error* (Technical Report No. 15). Palo Alto - CA (USA): Applied Mathematics and Statistics Laboratory, Stanford University.
- Hambleton, R. K. y Jones, R. W. (1993). Comparison of classical test theory and item response theory and their application to test development. *Educational Measurement: Issues and Practice*, 12(3), 38–47.
- Hambleton, R. K., Swaminathan, J. y Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage publications.
- Harvey, J. (1998). *Evaluation cookbook* (Informe Técnico). Learning Technology Dissemination Initiative (LTDI). Scottish Higher Education Funding Council.
- Hetter, R. D., Segall, D. O. y Bloxon, B. M. (1994). A comparison of item calibration media in computerized adaptive testing. *Applied Psychological Measurement*, 18(3), 197–204.
- Hively, W. (1974). Introduction to domain-reference testing. *Educational Technology*, 14(6), 5–10.
- Holt, P., Dubs, S., Jones, M. y Greer, J. (1994). The state of student modelling. En J. E. Greer y G. McCalla (Eds.), *Student modelling: The key to individualized knowledge-based instruction* (Vol. 125, pp. 3–35). New York: Springer Verlag.
- Hontangas, P., Ponsoda, V., Olea, J. y Abad, F. (2000). Los test adaptativos informatizados en la frontera del siglo xxi: una revisión. *Metodología de las Ciencias del Comportamiento*, 2(2), 183–216.
- Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicológica*, 21, 175–189.
- Härdle, W. (1992). *Applied nonparametric regression*. Cambridge: University Press.
- Huang, S. X. (1996a). A content-balanced adaptive testing algorithm for computer-based training systems. En C. Frasson, G. Gauthier y A. Lesgold (Eds.), *Lecture notes in computer science 1086. proceedings of the 3rd international conference on intelligent tutoring systems. its 1996* (pp. 306–314). New York: Springer Verlag.
- Huang, S. X. (1996b). On content-balanced adaptive adaptive testing. En A. D. de Ilarrazá Sánchez y I. F. de Castro (Eds.), *Lecture notes in computer science 1108. proceedings of the 3rd international conference on computer aided learning and instruction in science and engineering. calisce 1996* (pp. 60–68). New York: Springer Verlag.

- Hudson, K. (1999). Adaptive testing. *Windows and .NET Magazine Network.* <http://www.win2000mag.com>.
- Huff, K. L. y Sireci, S. G. (2001). Validity issues in computer based testing. *Educational Measurement: Issues and Practice*, 20(3), 16–25.
- Hulin, C. L., Drasgow, F. y Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow-Jones, Irwin.
- Hwang, G. J. (2003). A test-sheet-generation algorithm for multiple assessment requirement. *IEEE Transactions on Education*, 46(3), 329–337.
- IMS Global Learning Consortium. (2005, June). *Ims question and test interoperability*. <http://www.imsglobal.org/question/index.cfm>.
- Intralearn Software Corp. (2003, April). *Intralearn*. <http://www.intralearn.com>.
- Jettmar, E. y Nass, C. (2002). Adaptive testing: Effect on user performance. En *Proceedings of the sigchi conference on human factors in computing systems: Changing our world, changing ourselves* (pp. 129–134). ACM Press.
- Junker, B. W. (2000). On the interplay between nonparametric and parametric irt, with some thoughts about the future. En A. Boomsma, M. A. J. van Duijn y T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 247–276). New York: Springer Verlag.
- Junker, B. W. y Sijtsma, K. (2001a). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 258–272.
- Junker, B. W. y Sijtsma, K. (2001b). Nonparametric item response theory in action: An overview of the special issue. *Applied Psychological Measurement*, 25(3), 211–220.
- Kavcic, A., Privosnik, M., Marolt, M. y Divjak, S. (2002). Educational hypermedia system alice: an evaluation of adaptive features. En *Advances in multimedia, video and signal processing systems. wseas'02*.
- Kay, J. (1995). The um toolkit for cooperative user models. *User Modeling and User-Adapted Interaction*, 4(3), 149–196.
- Kay, J. (2000). Stereotypes, student models and scrutability. En G. Gauthier, C. Frasson y K. VanLehn (Eds.), *Lecture notes in computer science. proceedings of 3rd international conference on intelligent tutoring systems. its 2000* (pp. 19–30). Springer Verlag.
- Kingsbury, G. G. y Houser, R. L. (1993). Assessing the utility of item response models: Computerized adaptive testing. *Educational Measurement: Issue and Practice*, Spring, 21–27.
- Kingsbury, G. G. y Weiss, D. J. (1979). *An adaptive testing strategy for mastery decision* (Psychometric Method Program No. Research Report 79-5). Department of Psychology. University of Minnesota.
- Kingsbury, G. G. y Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2(4), 359–375.

- Klauer, K. C. y Sydow, H. (2000). Modeling learning in short-term learning tests. En A. Boomsma, M. A. J. van Duijn y T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 69–88). New York: Springer Verlag.
- Kommers, P., Grabinger, S. y Dunlap, J. (1996). *Hypermedia learning environments. instructional design and integration*. Lawrence Erlbaum Associates.
- Lawson, S. (1991). One parameter latent trait measurement: Do the result justify the effort? En B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (pp. 159–168). Greenwich, CT: JAI.
- Lazarsfeld, P. F. (1950). Studies in social psychology in world war ii. En (Vol. 4). Princeton University Press.
- Lilley, M. y Barker, T. (2004). A computer-adaptive test that facilitates the modification of previously entered responses: An empirical study. En J. C. Lester, R. M. Vicari y F. Paraguaçu (Eds.), *Proceedings of the 7th international conference on intelligent tutoring systems (its 2004). lecture notes in computer science* (pp. 22–33). New York: Springer Verlag.
- Lilley, M., Barker, T. y Britton, C. (2004). The development and evaluation of a software prototype for computer-adaptive testing. *Computers & Education*, 43, 109–123.
- Littman, D. y Soloway, E. (1988). Foundations of intelligent tutoring systems: An introduction. En M. C. Polson y J. J. Richardson (Eds.), *Evaluating its: The cognitive science perspective* (pp. 209–242). Lawrence Erlbaum Associates, Inc.
- Litwin, J. y Fernández, G. (1977). *Medidas, evaluación y estadísticas aplicadas a la educación física y el deporte*. Stadium.
- Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs*(7).
- Lord, F. M. (1970). Item characteristic curves estimated without knowledge of their mathematical form: A confrontation of birbaum's logistic model. *Psychometrika*, 35(43–50).
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Lord, F. M. y Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- López Cuadrado, J., Pérez, T. A., Arrubarrena, R. M., Vadillo, J. A. y Gutiérrez, J. (2002). Generation of computerized adaptive tests in an adaptive hypermedia system. En A. M. Vilas, J. A. M. González y I. S. de Zaldívar (Eds.), *Educational technology: International conference on tic's in education*.
- López Cuadrado, J., Pérez, T. A., Vadillo, J. A. y Arrubarrena, R. M. (2002). Integrating adaptive testing in an educational system. En A. M. Vilas, J. A. M. González y I. S. de Zaldívar (Eds.), *1st international conference on educational technology in cultural context*.
- Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

- Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20(4), 389–404.
- Machuca, E. L. y Guzmán, E. (2004). *Triviette, trivial por internet sobre el sistema siette*. Proyecto de fin de carrera, E.T.S.Ingeniería Informática. Dpto. Lenguajes y Ciencias de la Computación. Universidad de Málaga.
- Mackenzie, D. M. (1999). Recent developments in the tripartite interactive assessment delivery system (triads). En C. Eabry (Ed.), *Proceedings of the 3rd international computer assessment conference (caa). loughborough university*.
- Marcke, K. V. (1991). A generic task model for instruction. En *Proceedings of nato advanced research. workshop on instructional design models for computer based learning environments*.
- Mark, M. A. y Greer, J. E. (1993). Evaluation methodologies for intelligent tutoring systems. *International Journal of Artificial Intelligence in Education: Special Issue on Evaluation*, 4(2/3), 129–153.
- Martin, B. y Mitrovic, A. (2002). Automatic problem generation in constraint-based tutors. En S. A. Cerri, G. Gouardères y F. Paraguacu (Eds.), *Proceedings of the 6th international conference on intelligent tutoring systems (its 2002). lecture notes in computer science* (pp. 388–398). New York: Springer Verlag.
- Martin, J. y VanLehn, K. (1995). A bayesian approach to cognitive assessment. En P. Nichols, S. Chipman y R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 141–165). Hillsdale, NJ: Lawrence Erlbaum.
- Martínez Arias, R. (1995). *Psicometría: Teoría de los test psicológicos y educativos*. Madrid: Síntesis , S.A.
- Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 60, 523–547.
- Masters, G. N. y Wright, B. D. (1997). The partial credit model. En W. J. van der Linden y R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York: Springer Verlag.
- Mavrikis, M. y González Palomo, A. (2003). Mathematical, interactive exercise generation from static documents. En *Mathematical knowledge management symposium*.
- Mayo, M. y Mitrovic, A. (2001). Optimising its behaviour with bayesian networks and decision theory. *International Journal of Artificial Intelligence in Education*, 12, 124–153.
- Mazza, R. (2003). Using open student models in distance learning environments to help teachers provide adaptive tutoring. *Studies in Communication Sciences (SCOMS). Special Issue New Media in Education*, 12, 245–251.
- McCallon, E. L. y Schumacker, R. E. (2002). *Classical test analysis*. (ELM Metrics, Inc.)
- McCormack, C. y Jones, D. (1997). *Building a web-based education system*. Wiley Computer Publishing.

- Melis, E., Andres, E., Bündenbender, J., Frischauft, A., Goguadze, G., Libbrecht, P., Pollet, M. y Ullrich, C. (2001). Activemath: A generic and adaptive web-based learning environment. *International Journal of Artificial Intelligence in Education*, 12, 385–407.
- Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, 19(1), 91–100.
- Millán, E. (2000). *Sistema bayesiano para el modelado del alumno*. Tesis doctoral no publicada, Universidad de Málaga.
- Millán, E., García-Hervás, E., Guzmán, E., Rueda, A. y Pérez de la Cruz, J. L. (2003). Adaptation and generation in a web-based tutor for linear programming. En J. M. Cueva-Lovelle, B. M. González-Rodríguez, L. Joyanes-Aguilar, J. E. Labra-Gayo y M. P. Paul-Ruiz (Eds.), *Proceedings of the 3rd international conference on web engineering (icwe 2003). lecture notes in computer science* (pp. 124–127). New York: Springer-Verlag.
- Millán, E., García-Hervás, E., Guzmán, E., Rueda, A. y Pérez de la Cruz, J. L. (2004). Tapli: An adaptive web-based learning environment for linear programming. En R. Conejo, M. Urretavizcaya y J. L. P. de-la Cruz (Eds.), *Current topics in artificial intelligence. lecture notes in artificial intelligence* (pp. 676–687). New York: Springer-Verlag.
- Millán, E. y Pérez de la Cruz, J. L. (2002). Diagnosis algorithm for student modeling diagnosis and its evaluation. *User Modeling and User-adapted Interaction*, 12(2-3), 281–330.
- Millán, E., Pérez de la Cruz, J. L. y Suárez, E. (2000). An adaptive bayesian network for multilevel student modelling. En G. Gauthier, C. Frasson y K. VanLehn (Eds.), *Proceedings of 3rd international conference on intelligent tutoring systems. its'2000. lecture notes in computer science*. (pp. 534–543).
- Mills, G. N. y Steffen, M. (2000). The gre computer adaptive test: Operational issues. En W. J. van der Linden y C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75–100). Dordrecht (NL): Kluwer Academic Publishers.
- Minsky, M. (1986). *The society of mind*. New York: Simon and Schuster.
- Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, 51(2), 177–195.
- Mislevy, R. J. (1993). Introduction. En N. Frederiksen, R. J. Mislevy y I. I. Béjar (Eds.), *Test theory for a new generation of tests*. Lawrence Erlbaum Associates.
- Mislevy, R. J. y Almond, R. (1997). *Graphical models and computerized adaptive testing* (Informe Técnico No. 434). Center of the Study of Evaluation (CSE).
- Mislevy, R. J., Wingersky, M. S. y Sheehan, K. M. (1994). *Dealing with uncertainty about item parameters: Expected response functions* (Informe Técnico No. ETS Research Report 94-28-ONR). Education Testing Service.

- Mitrovic, A. (2002). Sint - a symbolic integration tutor. En S. A. Cerri, G. Gouardères y F. Paraguacu (Eds.), *Proceedings of the 6th international conference on intelligent tutoring systems (its 2002). lecture notes in computer science* (pp. 587–595). New York: Springer Verlag.
- Mitrovic, A. y Martin, B. (2002). Evaluating the effects of open student models on learning. En P. D. Bra, P. Brusilovsky y R. Conejo (Eds.), *Proceedings of the 2nd international conference on adaptive hypermedia and adaptive web-based systems. ah 2002. lecture notes in computer science* (pp. 296–305). New York: Springer Verlag.
- Mokken, R. J. (1997). Nonparametric models for dichotomous responses. En W. J. van der Linden y R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351–367). New York: Springer Verlag.
- Molenaar, I. W. (2001). Thirty years of nonparametric item response theory. *Applied Psychological Measurement*, 25(3), 295–299.
- Molnar, G. (2005). *Evaluación educativa*. http://www.chasque.net/gamolnar/evaluacion_educativa/homeevaluacion.html.
- Muñiz, J. (1997). *Introducción a la teoría de respuesta a los Ítems*. Madrid: Pirámide.
- Muñiz, J. (2002). *Teoría clásica de los tests*. Madrid: Pirámide.
- Muñiz, J. y Hambleton, R. K. (1999). Evaluación psicométrica de los tests informatizados. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests informatizados: Fundamentos y aplicaciones* (pp. 23–52). Pirámide.
- Muraki, E. (1990). Fitting a polytomous item response model to likert-type data. *Applied Psychological Measurement*, 14, 59–71.
- Muraki, E. (1992). A generalized partial credit model: Application to an em algorithm. *Applied Psychological Measurement*, 16, 159–176.
- Muraki, E. (1997). A generalized partial credit model. En W. J. van der Linden y R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). New York: Springer Verlag.
- Murray, T. (1993). Formative qualitative evaluation for 'exploratory' its research. *International Journal of Artificial Intelligence in Education: Special Issue on Evaluation*, 4(2/3), 179–207.
- Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement*, 15, 159–176.
- Neira, A., Brugos, J. A. y Alguero, A. (2002). Tests adaptativos informatizados utilizando conjuntos borrosos. *Revista de Metodología de las Ciencias del Comportamiento, Volumen especial*, 423–426.
- Nelson, L. R. (2003). *Some ctt and irt comments. lertap 5 documents series*. <http://www.lertap.curtin.edu.au>.
- Novell. (2004). *Novell testing theory web page*. <http://educational.novell.com/testinfo/theory.htm>.

- Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, 3, 1–18.
- Ohlsson, S. (1994). Constraint-based student modelling. En J. E. Greer y G. McCalla (Eds.), *Student modelling: The key to individualized knowledge-based instruction* (Vol. 125, pp. 167–190). New York: Springer Verlag.
- Olea, J. y Ponsoda, V. (2001). *Tests adaptativos informatizados* (Curso de doctorado). Programa de Metodología del Comportamiento.
- Olea, J., Ponsoda, V. y Prieto, G. (1999). *Tests informatizados: Fundamentos y aplicaciones*. Pirámide.
- Oppermann, R., Rashev, R. y Kinshuk. (1997). Adaptability and adaptivity in learning systems. *Knowledge Transfer*, II, 173–179.
- Osterlind, S. J. (1990). Towards a uniform definition of a test item. *Educational Research Quarterly*, 14(4), 2–5.
- Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance and other formats*. London: Kluwer Academic Publishers.
- Owen, R. J. (1969). *A bayesian approach to tailored testing* (Research Report No. 69-92). Educational Testing Service.
- Owen, R. J. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70(350), 351–371.
- Papanikolaou, K. A., Grigoriadou, M., Kornikolakis, H. y Magoulas, G. D. (2003). Personalizing the interaction in a web-based educational hypermedia system: the case of inspire. *User Modeling and User-Adapted Interaction*, 13, 213–267.
- Parshall, C. G., Davey, T. y Pashley, P. J. (2000). Innovate item types for computerized testing. En W. J. van der Linden y C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 129–148). Dordrecht (NL): Kluwer Academic Publishers.
- Patel, A., Kinshuk y Russell, D. (1998). A computer based intelligent assessment system for numeric disciplines. *Information Services and Use*, 18(1-2), 53–63.
- Patelis, T. (2000). *An overview of computer-based testing* (Research Notes). The College Board. Office of Research and Development.
- Pathak, S. y Brusilovsky, P. (2002). Assessing student programming knowledge with web-based dynamic parameterized quizzes. En P. Barker y S. Rebelsky (Eds.), *Proceedings of the ed-media 2002 - world conference on educational multimedia, hypermedia and telecommunications* (pp. 1548–1553).
- Pearl, J. (1988). *Probabilistic reasoning in expert systems: Networks of plausible inference*. San Francisco: Morgan Kaufmann Publishers, Inc.
- Pellegrino, J., Chudowsky, N. y Glaser, R. (2001). *Knowing what student knows: The science and design of educational assessment*. National Academy of Science.

- Piaget, J. (1952). *The origins of intelligence in children*. New York: International Universities Press.
- Pila Teleña, A. (1988). *Evaluación de la educación física y los deportes*. San José, C.R. : Olimpia.
- Planchard, E. (1986). *Orientaciones actuales de la pedagogía*. Buenos Aires: Troquel.
- Pérez, T. A., Gutiérrez, J. y Lopistéguy, P. (1995). An adaptive hypermedia system. En *Proceedings of the international conference of artificial intelligence in education. aied'95*.
- Pérez de la Cruz, J. L., Conejo, R. y Guzmán, E. (2005). Qualitative and quantitative student models. En C. K. Looi, G. McCalla, B. Bredeweg y J. Breuker (Eds.), *Artificial intelligence in education: Supporting learning through intelligent and socially informed technology. (aied 2005)* (pp. 531–538). Amsterdam: IOS Press.
- Promissor. (2003, April). *Catglobal*. <http://www.catglobal.com>.
- QuestionMark Corp. (2003, April). *Questionmark*. <http://www.questionmark.com>.
- Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, 56, 611–630.
- Ramsay, J. O. (2000). *Testgraf: A program for the graphical analysis of multiple choice test and questionnaire data [computer program]*. montreal: Mcgill university.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment test*. Danish Institute for Educational Research.
- Renom, J. y Doval, E. (1999). Tests informatizados: Fundamentos y aplicaciones. En J. Olea, V. Ponsoda y G. Prieto (Eds.), (pp. 63–83). Pirámide.
- Revuelta, J. (2000). *A psychometric model for multiple choice items*. Tesis doctoral no publicada, Universidad Autónoma de Madrid.
- Revuelta, J., Ponsoda, V. y Olea, J. (1998). Métodos para el control de las tasas de exposición en tests adaptativos informatizados. *Revista Electrónica de Investigación y Evaluación Educativa (RELIEVE)*. <http://www.uv.es/RELIEVE/v4n2/RELIEVEv4n2.htm>, 4(2).
- Reye, J. (2002). A belief net backbone for student modelling. En S. A. Cerri, G. Gouardères y F. Paraguacu (Eds.), *Proceedings of the 6th international conference on intelligent tutoring systems (its 2002). lecture notes in computer science* (p. 596-604). New York: Springer Verlag.
- Rocklin, T. R. y O'Donnell, A. M. (1987). Self-adaptive testing: A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology*, 79, 315–319.
- Rodríguez Artacho, M. (2000). *Una arquitectura cognitiva para el diseño en entornos telemáticos de enseñanza y aprendizaje*. Tesis doctoral no publicada, Escuela Técnica Superior de Ingenieros Industriales. Universidad Nacional de Educación a Distancia (UNED).

- Román, F. y Conejo, R. (2003). *Editor de escenas para siette*. Proyecto de fin de carrera, E.T.S.Ingeniería Informática. Dpto. Lenguajes y Ciencias de la Computación. Universidad de Málaga.
- Ríos, A., Conejo, R., Trella, M., Millán, E. y Pérez de la Cruz, J. L. (1999). Aprendizaje automático de las curvas características de las preguntas en un sistema de generación automática de tests. En A. García, R. Rizo, S. Moral y F. Toledo (Eds.), *Actas de la conferencia española para la inteligencia artificial. caepia'99*. Murcia.
- Ríos, A., Millán, E., Trella, M., Pérez de la Cruz, J. L. y Conejo, R. (1998). Internet based evaluation system. En S. Lajoie y M. Vivet (Eds.), *Open learning environments: New computational technologies to support learning, exploration and collaboration. proceedings of the 9th world conference of artificial intelligence and education aied'99* (pp. 387–394). San Antonio (Texas): IOS Press.
- Ríos, A., Pérez de la Cruz, J. L. y Conejo, R. (1998). Siette: Intelligent evaluation system using test for teleeducation. En *Workshop on intelligent tutoring systems on the web. 4th international conference on intelligent tutoring system. its'98*. San Antonio (Texas).
- Roskam, E. E. (1997). Models for speed and time-limit tests. En W. J. van der Linden y R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 187–208). New York: Springer Verlag.
- Rudner, L. M. (2002). An examination of decision-theory adaptive testing procedures. En *Annual meeting of the american educational research association*.
- Ruiz de Pinto, L. (2002). Evaluación. tipos de evaluación. *Revista de Postgrado de la VI Cátedra de Medicina*, 118.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, 17.
- Samejima, F. (1995). A cognitive diagnosis method using latent trait models: Competency space approach and its relation with dibello and stout's unified cognitive-psychometric diagnosis model. En P. Nichols, S. Chipman y R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 391–410). Hillsdale, NJ: Lawrence Erlbaum.
- Samejima, F. (1997). The graded response model. En W. J. van der Linden y R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer Verlag.
- Samejima, F. (1998). Efficient nonparametric approaches for estimating the operating characteristic of discrete item responses. *Psychometrika*, 63, 111–130.
- Sands, W. A., Waters, B. K. y McBride, J. R. (1996). *Computerized adaptive testing: From inquiry to operation*. Washington, DC (USA): American Psychological Association.
- Santisteban, C. (1990). *Psicometría: Teoría y práctica en la construcción de tests*. Las Rozas (Madrid): Norma.
- Santisteban, C. y Alvarado, J. (2001). *Modelos psicométricos*. Madrid: UNED.
- Schank, R. C. y Cleary, C. (1994). *Engines for education*. Lawrence Erlbaum Associates.

- Scriven, M. (1967). The methodology of evaluation. En R. E. Stake (Ed.), *Perspectives of curriculum evaluation*. Chicago, IL: Rand McNally.
- Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61(1), 331–354.
- Segall, D. O. (2001). General ability measurement: An application of multidimensional item response theory. *Psychometrika*, 66(1), 79–97.
- Segall, D. O. y Moreno, H. E. (1999). Development of the computerized adaptive testing version of the armed service vocational aptitude battery. En F. Drasgow y J. B. Olson-Buchanan (Eds.), *Innovations in computer assisted assessment*. Mahwah, NJ (USA): Lawrence Erlbaum Associates.
- Self, J. A. (1994). Formal approaches to student modeling. En J. E. Greer y G. McCalla (Eds.), *Student modeling: The key to individualized knowledge-based instruction* (Vol. 125, pp. 295–352). New York: Springer Verlag.
- Shavelson, R. J. y Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, NJ: Sage Publications.
- Shute, V. J. y Regian, J. W. (1993). Principles for evaluating intelligent tutoring systems. *International Journal of Artificial Intelligence in Education: Special Issue on Evaluation*, 4(2/3), 245–271.
- Sijtsma, K. y Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks: Sage Publications.
- Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer-Verlag.
- Skinner, B. F. (1985). *Aprendizaje y comportamiento*. Martínez-Roca.
- Sleeman, D. y Brown, J. S. (1982). *Intelligent tutoring systems*. Academic Press, Inc.
- Smits, D. J. M., De Boeck, P. y Hoskens, M. (2003). Examining the structure of concepts: Using interactions between items. *Applied Psychological Measurement*, 27(6), 415–439.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15, 72–101.
- Stage, C. (1998). A comparison between item analysis based on item response theory and classical test theory: a study of the swesat subtest read. *Educational Measurement*, 30.
- Stauffer, K. (1996). *Student modeling and web-based learning systems*. <http://ccism.pc.athabascau.ca/html/students/stupage/project/abstract.html>.
- Stout, W. (2001). Nonparametric item response theory: A maturity and applicable measurement modeling approach. *Applied Psychological Measurement*, 25(3), 300–306.
- Tam, S. S. (1992). *A comparison of methods for adaptive estimation of a multidimensional trait*. Tesis doctoral no publicada, Graduate School of Arts and Science, Columbia University. Order number: 9221219.

- Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, 12(1), 55–73.
- Thissen, D. (1988). *Multilog: Multiple, categorical item analysis and test scoring using item response theory (versión 5.1)*. Mooresville, IN: Scientific Software.
- Thissen, D. (1993). Repealing rules that no longer apply to psychological measurement. En N. Frederiksen, R. J. Mislevy y I. Béjar (Eds.), *Test theory for a new generation of tests*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Thissen, D. y Mislevy, R. (1990). Testing algorithms. En H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 103–136). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.
- Thissen, D. y Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, 49, 501–519.
- Thissen, D. y Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, 51, 567–577.
- Thissen, D. y Steinberg, L. (1997). A response model for multiple choice items. En W. J. van der Linden y R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51–65). New York: Springer Verlag.
- Thorndike, R. L. (1984). Educational measurement: Theory and practice. En D. Spearritt (Ed.), *The improvement of measurement in education and psychology: Contributions of latent trait theory* (pp. 2–13).
- Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 16, 433–451.
- Trella, M., Conejo, R., Bueno, D. y Guzmán, E. (2002). An autonomous component architecture to develop www-its. En *Proceedings of the workshop on adaptive systems for web-based education. ah 2002* (pp. 69–80).
- Trella, M., Conejo, R., Guzmán, E. y Bueno, D. (2003). An autonomous component architecture to develop www-its. En *Proceedings of the 3rd international conference on web engineering (icwe 2003). lecture notes in computer science*. New York: Springer Verlag.
- Twidale, M. (1993). Redressing the balance: The advantages of informal evaluation techniques for intelligent learning environments. *International Journal of Artificial Intelligence in Education: Special Issue on Evaluation*, 4(2/3), 155–178.
- Urretavizcaya, M. (2001). Sistemas inteligentes en el ámbito de la educación. *Inteligencia Artificial. Revista Iberoamericana De Inteligencia Artificial*, 12, 5–12.
- van der Linden, W. J. y Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Kluwer Academic Publishers.
- van der Linden, W. J. y Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer Verlag.
- van der Linden, W. J. y Pashley, P. J. (2001). Item selection and ability estimation in adaptive testing. En W. J. van der Linden y C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–26). Kluwer Academic Publisher.

- VanLehn, K. (1988). Student modeling. En M. C. Polson y J. J. Richardson (Eds.), *Foundations of intelligent tutoring systems* (pp. 55–76). Lawrence Erlbaum Associates Publishers.
- VanLehn, K. y Martin, J. (1998). Evaluation of an assessment system based on bayesian student modeling. *International Journal of Artificial Intelligence and Education*, 8(2).
- VanLehn, K., Niu, Z., Siler, S. y Gertner, A. S. (1998). Student modeling from conventional test data: A bayesian approach without priors. En B. Goettl, C. L. Redfield, H. M. Halff y V. J. Shute (Eds.), *Proceedings of 4th international conference on intelligent tutoring systems. its'98. lecture notes in computer science* (Vol. 1452, pp. 434–443).
- Vassileva, J. (1997). Dynamic course generation on the www. En B. du Bolay y R. Mizoguchi (Eds.), *Knowledge and media in learning systems. proceedings of the 8th world conference on artificial intelligence in education aied'97* (pp. 498–505).
- Verdejo, M. F. (1994). Building a student model for an intelligent tutoring system. En J. E. Greer y G. McCalla (Eds.), *Student modelling: The key to individualized knowledge-based instruction* (Vol. 125, pp. 147–163). New York: Springer Verlag.
- Verhelst, N. D., Verstralen, H. H. F. M. y Jansen, M. G. H. (1997). A logistic model for time-limit tests. En W. J. van der Linden y R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 169–186). New York: Springer Verlag.
- Verstralen, H. H. F. M. (1997a). *A latent irt model for options of multiple choice items* (Measurement and Research Department Reports No. 97-5). Cito.
- Verstralen, H. H. F. M. (1997b). *A logistic latent class model for multiple choice items* (Measurement and Research Department Reports No. 97-1). Cito.
- Vygotskii, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wainer, H. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.
- Wainer, H. (2000). Cats: Whither and whence. *Psicológica*, 21, 121–133.
- Wainer, H. y Mislevy, R. J. (1990). Item response theory, item calibration and proficiency estimation. En H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 65–102). Lawrence Erlbaum Associates, Inc.
- Wainer, H., Morgan, A. y Gustafsson, J. E. (1980). A review of estimation procedures for the rasch model with an eye toward logish tests. *Journal of Educational Statistics*, 5, 35–64.
- Wand, M. P. y Jones, M. C. (1995). *Kernel smoothing*. Chapman and Hall.
- Wang, W. C. y Chen, P. H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, 28(5), 295–316.
- WBT Systems. (2003, April). *Topclass*. <http://topclass.uneg.edu/>.
- WebCT Inc. (2003, April). *Webct*. <http://www.webct.com>.

- Weber, G. (1996). Individual selection of examples in an intelligent learning environment. *Journal of Artificial Intelligence in Education*, 7(1), 3–31.
- Weber, G. y Brusilovsky, P. (2001). Elm-art: An adaptive versatile system for web-based instruction. *International Journal of Artificial Intelligence in Education*, 12, 351–383.
- Weber, G., Kuhl, H. C. y Weibelzahl, S. (2001). Developing adaptive internet based courses with the authoring system netcoach. En *Proceedings of the 3rd workshop on adaptive hypertext and hypermedia. sonthofen (germany)* (pp. 35–48).
- Weber, G. y Specht, M. (1997). User modeling and adaptive navigation support in www-based tutoring systems. En A. Jameson, C. Paris y C. Tasso (Eds.), *Proceedings of the 6th international conference on user modelling um'97* (pp. 289–300).
- Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473–492.
- Welch, R. E. y Frick, T. W. (1993). Computerized adaptive testing in instruction settings. *Educational Technology Research and Development*, 41(3), 47–62.
- Wenger, E. (1987). *Artificial intelligence and tutoring systems*. Morgan Kaufmann Publishers, Inc.
- White. (1982). Some major components in general intelligence. En H. J. Eysenck (Ed.), *A model for intelligence*. New York: Springer Verlag.
- Winne, P. H. (1993). A landscape of issues in evaluating adaptive learning systems. *International Journal of Artificial Intelligence in Education: Special Issue on Evaluation*, 4(2/3), 273–332.
- Wise, S. L. (1999). Tests autoadaptados informatizados: fundamentos, resultados de investigación e implicaciones para la aplicación práctica. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests informatizados: Fundamentos y aplicaciones* (pp. 189–203). Pirámide.
- Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8, 338–353.

Índice alfabético

- τ -equivalencia, 24
ítems, 6, 10, 15, 20, 169, 191
 siettlets, 172, 176, 177, 195, 207, 209,
 210, 305
 adivinanza, 29, 92
 autocorregidos, 172
 banco de ítems, 21, 42, 132, 137, 218
 calibración, 9–11, 19, 49, 64, 83, 89,
 93, 113, 122, 123, 161, 199, 244,
 303
 en línea, 54
 inicial, 54
 Curva Característica del Ítem, 26, 40,
 115, 218
 curvas características, 8
 Curvas Características de Opción, 116
 Curvas Características de Respuesta,
 33, 56, 116
 de asociación, 121, 153, 219
 de completar, 22
 de correspondencia, 21
 de emparejamiento, 120, 153, 219, 220
 de opción múltiple, 21, 117, 152, 169,
 218, 220
 de ordenación, 21, 119, 153, 169, 219,
 220
 de redacción, 22
 de relación, 120, 169
 de respuesta corta, 18, 21, 169, 207
 de respuesta múltiple, 21, 118
 con opciones dependientes, 119, 153,
 169, 219, 220
 con opciones independientes, 118,
 153, 169, 219
 de respuesta sobre figura, 21, 172, 176
 de verdadero/falso, 21, 117, 152, 169,
 218, 220
 dificultad, 25, 28, 39, 52, 76, 95, 156
 discriminación, 25, 28, 36–39, 88, 95
 distractores, 34, 176
externos, 182
facilidad, 52
generación automática de ítems, 174
 basada en Gramáticas de Contexto
 Libre, 174
 basada en mecanismos simples, 174
 basada en técnicas ad-hoc, 174
generativos, 174, 181
modelado de ítems, 174
por partes, 22, 33
respuesta latente, 34
sobreexposición de los ítems, 174
temporizados, 182
ítems o test de anclaje, 50
, 30
ActiveMath, 79, 102
ALICE, 93, 103
alumno, 20
alumnos simulados, 216
anclaje, 50
ANDES, 96
AnimalWatch, 172
Asociación de un ítem a un concepto, 137
ayudas, 187
biblioteca de errores, 72
bootstrapping, 122, 171, 303
C-Quest, 60
CALEAP-WEB, 103
calibración, 10, 114
CARAT, 48
CATGlobal, 61
CBAT-2, 83, 89–91, 93, 98, 102
CLARISSE, 82
clusterización, 95
coeficiente de Cronbach, 25
Computerized Adaptive Testing version of
the Armed Services Vocational
Aptitude Battery, 20

- concepto, 135
 Conductivismo, 66
 Constructivismo, 69, 306
 contribución de una alternativa a la información del ítem, 57
 corrección por atenuación, 24
 crédito parcial, 32, 36
 Criterio de evaluación por puntos, 152, 194
 Criterio de evaluación por puntos con penalización, 194
 Criterio de evaluación porcentual, 152, 194
 Criterio de finalización basado en el tiempo límite, 48, 161, 194
 Criterio de finalización basado en la máxima precisión alcanzable, 160
 Criterio de finalización basado en la precisión de la estimación, 160
 Criterio de finalización basado en la probabilidad mínima, 159
 Criterio de finalización de longitud fija, 48, 161, 194
 Criterio de finalización de longitud variable, 48
 criterio de finalización de un test, 43, 48, 58, 159, 194
 Criterio de selección aleatoria, 193
 Criterio de selección basado en el nivel de dificultad, 47, 155
 Criterio de selección basado en la entropía, 157
 Criterio de selección basado en la máxima información, 46, 57, 88, 158
 Criterio de selección bayesiano de la máxima precisión esperada, 47, 154
 criterio de selección de ítems, 43, 46, 57, 84, 154, 193
 Criterio de selección según dificultad ascendente, 193
 Criterio de selección según orden ascendente, 193
 cuadratura gaussiana, 51
 currículo, 68, 135, 189
- DCG, 78, 102
 diagnóstico del alumno, 8, 69, 74, 112, 132, 141, 146, 302, 304
- Edie, 192
 Educational Testing Service, 20, 106, 124
- ELM-ART, 75, 81, 102
 ELM-PE, 76
 entropía, 87, 157
 entropía esperada, 157
 equiparación, 50
 error cuadrático medio, 127, 245
 error de medida, 22
 escalas de tipo Likert, 32, 63
 Esperanza/Maximización, 51, 97
 estereotipos, 82
 evaluación, 3
 educativa, 3, 4, 8
 (con referencia) a un criterio, 6
 formativa, 5, 199, 205
 normativa, 5
 predictiva, 5, 199
 sumativa, 5, 199, 205
 formativa, 215, 271
 sumativa, 216, 217
 Evaluación agregada, 148
 Evaluación completa, 148
 Evaluación completa con retropropagación, 149
 Evaluación de un ítem sobre un concepto, 139
 Evaluación de un test sobre un concepto, 143
 Evaluación directa de un ítem sobre un concepto, 138
 Evaluación directa de un test sobre un concepto, 141
 Evaluación indirecta ascendente (o hacia arriba) de un test sobre un concepto, 142
 Evaluación indirecta de un ítem sobre un concepto, 138
 Evaluación indirecta de un test sobre un concepto, 142
 Evaluación indirecta descendente (o hacia abajo) de un test sobre un concepto, 142
 examen, 4
- fórmula de Spearman-Brown, 25
 FastTEST, 61
 función característica del test, 49
 Función de evaluación de la respuesta, 114
 función de información, 46, 57, 90
 función de información de la alternativa, 57

- Función de Respuesta Esperada, 177
Función respuesta seleccionada, 114
- Graduate Record Examination, 20
GTE, 78
- Hermes, 192
HEZINET, 85, 103
Hypertutor, 81
- I-assess, 60
INSPIRE, 88, 93, 102, 103
Inteligencia Artificial, 7, 9, 62, 66, 74
Interbook, 75
Intralearn, 59
ITED, 95
ITEMSYS, 61
- Java EE, 207
JESS, 79
- lógica difusa, 91, 103
LeActiveMath, 211, 306
lenguajes de autor, 66
- Método de calibración basado en el suavizado núcleo, 54
Método de calibración basado en la máxima verosimilitud condicional, 52
Método de calibración basado en la máxima verosimilitud conjunta, 51
Método de calibración basado en la máxima verosimilitud marginal, 51
Método de calibración bayesiano, 53
Método de estimación basado en la máxima verosimilitud, 45, 57, 58, 127
Método de estimación bayesiano, 45
Método de estimación bayesiano de la Esperanza a Posteriori, 46, 152, 194, 220
Método de estimación bayesiano del Máximo a Posteriori, 46, 152, 160, 194, 220
método del Centro de Gravedad, 94
módulo de adaptación, 133
módulo de evaluación, 133
módulo de presentación, 133
MEDEA, 204, 211, 306
METSAS, 95
MicroCAT, 60
modelo conceptual, 132, 135
- Modelo de crédito parcial generalizado, 36, 38
dificultad del paso, 37
intersección de categorías, 37
modelo de evaluación cognitiva, 10, 131
Modelo de Homogeneidad Monótona, 40
modelo de instrucción, 69
Modelo de Monotonicidad Doble, 40
modelo de respuesta, 42
modelo de respuesta cuasipolítómico o parcialmente polítómico, 122
Modelo de respuesta graduada, 32, 35
Curvas Características de Funcionamiento, 35
Curvas de Respuesta Categóricas, 36
Modelo de respuesta nominal, 36
parámetro de intercepción, 36
Modelo de respuesta para ítems de opción múltiple de Thissen y Steinberg, 37
modelo de tareas, 21
modelo del alumno, 17, 68, 70, 133, 145
abierto, 73
basado en intervalos de confianza, 72
basado en lógica difusa, 73
basado en restricciones, 72
de perturbación, 72
de recubrimiento o superposición, 71
diferencial, 71
inspeccionables y/o modificables, 74
modelo del dominio, 67, 76, 132, 134
abierto, 76
basado la metodología de los sistemas expertos, 68
cognitivo, 68
de caja negra, 68
Modelo Granularidad-Bayes, 98
modelo pedagógico, 69
Modelo-Vista-Controlador, 207
Modelos basados en el proceso de respuesta de Nedelsky, 37
Full Nedelsky, 38
Modelo de Revuelta para ítems de opción múltiple, 39
MULTILOG, 32
- NetCoach, 76
Newton-Raphson, 45, 53, 112
nivel de conocimiento, 8, 15, 42, 44

- OLAE, 96
- PARDUX, 61
- PASS, 89, 102
- pedagogía por objetivos, 6
- planificador de instrucción, 69
- POLA, 98
- probabilidad marginal, 51
- programación dinámica, 95
- Programación Lineal, 203
- programas lineales, 66
- programas ramificados, 66
- psicometría, 15
- PTS, 95
- puntuación en el test, 23
- puntuación verdadera, 23, 24
- QTI, 197
- QUANTI, 81, 103
- QuestionMark, 59
- QUIZPACK, 175
- reconocimiento de patrones, 66
- red neuronal ART, 95
- red semántica, 68, 135
- redes bayesianas, 73, 95, 104
- refuerzo, 17, 71, 73, 76, 81, 88, 90, 167, 171, 187, 188, 191, 192, 306, 308
- relación de inclusión, 135
- Relación de inclusión directa entre dos conceptos, 135
- Relación de inclusión entre dos conceptos, 136
- Relación de inclusión indirecta entre dos conceptos, 135
- orden de la relación, 136
- S-QTI, 197
- Scholastic Aptitude Test, 16
- servicios Web, 204
- sesión, 21
- SIETTE, 10, 12, 167, 217, 266, 268, 272, 273, 305
- ítems básicos, 169
- analizador de resultados, 197, 206
- aula virtual, 186, 206
- base de conocimientos, 184, 206
- biblioteca de plantillas, 177
- ejercicios de completar, 178
- ejercicios de emparejamiento, 179
- ejercicios de inserción en un conjunto, 178
- ejercicios de ordenación, 178
- ejercicios de pasatiempos, 179
- calibrador de ítems, 206
- editor de tests, 189, 206
- repositorio de modelos del alumno, 184, 206
- repositorio de profesores, 185, 206
- simulador, 217
- Sistema de Evaluación Inteligente Bizantino, 174
- sistemas adaptativos, 66
- Sistemas de Enseñanza Asistida por Ordenador, 65
- Sistemas Educativos Adaptativos para la Web, 70, 301
- sistemas generativos, 66
- sistemas hipermedia, 69
- Sistemas Tutores Inteligentes, 8, 9, 11, 62, 65, 66, 102, 132, 146, 199, 301
- suavizado núcleo, 54, 64, 124, 245
- función núcleo, 56, 57, 127
- parámetro de suavizado, 56, 127, 245
- TANGOW, 80, 102
- TAPLI, 203, 211, 306
- TEA, 203
- Teoría Clásica de los Tests, 23, 167, 177
- Modelo de tests τ -congenéricos, 25
- Modelo de tests τ -equivalentes esenciales, 25
- Modelo de tests paralelos, 24
- Teoría de la Generalizabilidad, 25
- Teoría Clásica de los Tests), 23
- Teoría de la Decisión, 86, 102
- Teoría de la Información, 87, 157
- Teoría de la Medida, 22
- Teoría de Respuesta al Ítem, 8, 10, 11, 19, 23, 26, 44, 131, 302
- modelos binarios, 30
- modelos con parte izquierda añadida, 31
- modelos con parte izquierda añadida y división por el total, 32
- modelos de división por el total, 31
- modelos diferenciales, 31
- modelos logísticos, 28
- de dos parámetros, 28, 39, 53

- de tres parámetros, 28, 53, 85, 86, 89, 101, 218
de un parámetro, 28, 53
modelos multidimensionales, 27, 306
modelos no paramétricos, 10, 30, 39, 303
modelos normales, 28
modelos paramétricos, 30
modelos polítómicos, 10, 11, 30, 32, 40, 57, 302
ordinales, 33
modelos unidimensionales, 27
Teoría de Respuesta al Ítem
modelos dicotómicos, 339
Teoría del Espacio de Conocimiento, 91
Teoría del Muestreo, 22
teorías de los tests, 11, 22
TerraNova CAT, 61
Test of English as a Foreign Language, 20
TESTGRAF, 55, 56
tests, 4, 6, 8, 15, 20, 41, 132, 141, 192
testlets, 202
Auto-adaptados Informatizados, 19
con referencia a un criterio, 19
de aptitud, 16
de autoevaluación, 17, 186
de habilidad, 16
de logros, 16
de papel y lápiz, 15, 43, 54, 169, 217, 266–268, 270, 306
de personalidad, 16
de ramificación, 46, 91
con estructura en árbol (o pirámides), 46
en dos estados, 46
fija, 46
fiabilidad, 7
independencia local, 27, 40, 113
invarianza, 27, 40
lineales, 18
monotonicidad, 40, 113
objetividad, 7
post-tests, 17
pretests, 17, 75, 199
puntuación verdadera, 270
temporizados, 195
testlets, 18, 118, 188, 274
unidimensionalidad, 27, 40, 113
validez, 7
Tests Adaptativos Bayesianos, 100, 102
Tests Adaptativos Informatizados, 8–11, 19, 41, 54, 83, 103, 132, 146
Tests Administrados por Computador, 17, 41
tests con multidimensionalidad entre ítems, 306
tests multidimensionales intra-ítems, 306
TopClass, 59
TRIADS, 178
TRIVIETTE, 204, 306
Webassessor, 60
WebCT, 59