ARTÍCULO

Estudio de un mecanismo para la administración adaptativa de ayudas en la realización de tests*

Ricardo Conejo, Eduardo Guzmán, José-Luis Pérez-de-la-Cruz, Eva Millán

Departamento de Lenguajes y Ciencias de la Computación.
Universidad de Málaga.
Bulevar Louis Pasteur, 35.
Málaga, 29071.
{conejo,guzman,perez,eva}@lcc.uma.es

Resumen

En este artículo se presenta una propuesta para la introducción de mecanismos adaptativos para la elección de ayudas, en un entorno de evaluación basado en tests adaptativos. El artículo comienza con una discusión sobre aspectos relativos a mecanismos de selección adaptativos de pistas, que resultan en la definición de dos axiomas que tales pistas deben cumplir. Posteriormente, se presenta una evaluación empírica preliminar con estudiantes reales, cuyo objetivo es evaluar un banco inicial de preguntas con sus respectivas pistas, para determinar si éstas son realmente útiles. La evaluación se llevará a cabo a partir de los resultados obtenidos tras administrar un test a muestra poblacional formada por alumnos reales con diferentes niveles de conocimiento.

Palabras clave: Administración adaptativa de ayudas, Tests Adaptativos Informatizados, Teoría de Respuesta al Ítem.

1. Introducción

La evaluación mediante pruebas tipo test se usa de modo habitual en diferentes contextos educativos y con diversos objetivos: calificación, autoevaluación, diagnóstico cognitivo, etc. Para mejorar la eficiencia del proceso de diagnóstico, los sistemas de tests adaptativos seleccionan la mejor pregunta de acuerdo a la estimación actual de características relevantes del examinando. De este modo, se alcanza un grado de exactitud mayor en la evaluación, a la vez que se reduce significativamente el número de preguntas. En este sentido, se pueden encontrar diferentes propuestas para tests adaptativos [12], [3]. Una de las teorías más sólidas utilizadas como fundamento en este tipo de tests es la *Teoría de la Respuesta al Item* (TRI) [9], la cual supone que la respuesta a una pregunta depende de un rasgo latente desconocido. Este rasgo latente θ representa un factor psicológico que se desea medir a través del test, y que no es directamente observable. En entornos educativos, se corresponde con el conocimiento del alumno que se examina.

En cualquier entorno educativo adaptativo es necesario tener estimaciones fiables del nivel de

*Este trabajo ha sido desarrollado en el marco del proyecto TIC203-04480, subvencionado por el Ministerio de Ciencia y Tecnología español.

conocimiento del alumno, con objeto de elegir qué tipo de acción instructiva es más apropiada en cada momento. En este sentido, los *Tests Adaptativos Informatizados* (TAI) [15] basados en la TRI proporcionan una herramienta de diagnóstico fiable y eficiente. SIETTE (*Sistema Inteligente de Evaluación mediante Tests*) [4], [8] es un entorno de evaluación a través de Internet, que permite construir y administrar tests convencionales y TAI, basados en una discretización de la TRI. Una de las características principales de SIETTE es que es una herramienta que puede ser integrada sin dificultad en cualquier entorno de aprendizaje Web (para un ejemplo de integración, véase [11]). Al realizar esta integración, SIETTE gestiona todo lo relativo al modelado del alumno (básicamente, creación y mantenimiento del modelo del alumno). Este sistema está accesible en la siguiente dirección: *http://www.lcc.uma.es/SIETTE*.

Por otro lado, actualmente no cabe duda de que una de las contribuciones principales a la psicología educativa en el siglo XX es la *Zona del Desarrollo Próximo* (ZDP) de Vigotsky [14]. Una definición breve y operativa para nuestros objetivos actuales es la dada en [16]: *la zona definida mediante la diferencia entre el rendimiento que demuestra un niño* (en nuestro caso, una persona) *en un test en dos condiciones diferentes: con y sin ayuda*.

Poco después de la definición de la ZDP, surgen los primeros intentos de aplicar este concepto en el contexto de la administración de tests, bajo las dos condiciones descritas (con o sin ayuda). Normalmente el objetivo es clasificar a los alumnos para poder asignarlos a los programas educativos más apropiados. El principal objetivo del trabajo que presentamos aquí es diferente: construir un modelo que permita la integración de mecanismos adaptativos de selección de ayudas en el sistema SIETTE.

Las pistas constituyen una táctica adecuada y eficiente en el proceso de enseñanza. Muchos sistemas tutores inteligentes (STI) también dan pistas al alumno, como por ejemplo ANDES [7], que evalúa a los alumnos teniendo en cuenta, no sólo el número de respuestas correctas, sino también el número de pistas recibidas; o AnimalWatch [1], que tiene diferentes tipos de pistas disponibles (muy/poco interactivas, concretas/simbólicas) y las selecciona adaptativamente en base a rasgos del usuario, tales como el nivel de desarrollo cognitivo y el sexo. Asimismo, estudios psicológicos demuestran que un docente suele usar una estimación aproximada del rendimiento del alumno para seleccionar la pista más apropiada [10].

Consecuentemente, supondremos que la ayuda se representa mediante pistas que denotaremos h_1, ..., h_n, que proporcionan diferentes niveles de ayuda para cada pregunta tipo test (a las que denominaremos, en adelante, *ítems*). Entenderemos que el mecanismo de selección de ayudas es adaptativo cuando la pista sea seleccionada automáticamente por el sistema. La filosofía que subyace en esta adaptación es que se debe tener en cuenta la distancia a la que se encuentra la pista en la ZDP, de modo que se seleccionará siempre aquella pista que proporcione la información mínima que necesita el alumno para ser capaz de responder correctamente al ítem.

El trabajo que se presenta en este artículo extiende de nuestras investigaciones previas sobre la introducción de pistas y mensajes de refuerzo en tests adaptativos [5]. Ahora nuestros objetivos son:

- La definición de un marco teórico para la selección adaptativa de pistas.
- La realización de un estudio empírico que permita validar la propuesta. Con este objetivo, hemos desarrollado un banco de ítems real para una asignatura, con sus respectivas pistas asociadas. Este banco ha sido usado en varios experimentos con examinandos reales, con el objetivo de validar el banco de pistas.

Una cuestión que merece la pena considerar, y que también es objeto de estudio en este artículo, es la calibración de los pares *ítem+pista*. Una vez llevado a cabo el procedimiento de calibración, se podrán administrar tests en los que tanto los ítems como las pistas se seleccionen de modo adaptativo, de acuerdo a la estimación actual del nivel de conocimiento del alumno.

La estructura del resto del artículo es la siguiente: en la sección 2 se introducen las nociones básicas sobre la teoría de los TAI y la TRI. En la sección 3 discutimos diversos aspectos relativos a la introducción de pistas en entornos de tests adaptativos. La sección 4 presenta los resultados de los experimentos realizados, en los que se ha desarrollado un banco preliminar de ítems con sus correspondientes pistas que, posteriormente, han sido presentados a tres grupos de alumnos reales. El objetivo de estos experimentos era determinar, de modo empírico, la validez (e idoneidad)

de estas pistas. El artículo termina presentando las conclusiones que hemos alcanzado en este estudio, junto con algunas líneas de trabajo futuro.

2. Los Tests Adaptativos Informatizados y la Teoría de Respuesta al Ítem

Los TAI y la TRI constituyen hoy en día el marco teórico estándar para una evaluación fiable y eficiente del alumno. Un *test adaptativo informatizado* se define como aquella prueba de evaluación en la cual la decisión de qué ítem presentar en cada momento y la decisión de cuando finalizar el test se adoptan, de un modo dinámico, en función del rendimiento del alumno en los ítems anteriores. Si se comparan los TAI realizados por dos examinandos diferentes, normalmente cada uno de ellos recibe ítems diferentes y en distinto orden. Para poder administrar TAI de forma adecuada, a cada ítem i de un test se le asigna la llamada *Curva Característica del Ítem* (CCI). Ésta es una función que representa la probabilidad de que el alumno dé una respuesta correcta a ese ítem, dado su nivel de conocimiento θ, que es el rasgo latente no observable que deseamos estimar a través de la realización del test. Una hipótesis importante es que este rasgo latente no cambia durante el test, es decir, que el alumno no aprende durante el test. La probabilidad P_i de contestar correctamente a un ítem ($u_i = 1$) se puede calcular mediante la expresión $P_i = P(u_i = 1|\theta)$, y, en consecuencia, la probabilidad Q_i de fallar ($u_i = 0$) vendrá dada por la expresión $Q_i = P(u_i = 0|\theta) = 1 - P(ui = 1|\theta)$. Si el test se compone de n ítems, conocemos las CCI, y suponemos la independencia local de los ítems, la función de verosimilitud L se calcula mediante la expresión:

$$L(u_1, u_2, ..., u_n|\theta) = \prod_{i=1}^{n} P_i^{u_i} Q_i^{1-u_i} \qquad (1)$$

El máximo de esta función da una estimación del valor más probable de θ. La distribución de probabilidad de θ se puede obtener fácilmente aplicando n veces la regla de Bayes. Una hipótesis habitual es que las CCI pertenecen a una familia de funciones que dependen de uno, dos o tres parámetros, y que se construyen basándose en distribuciones normales o logísticas. Así, por ejemplo, el modelo logístico de tres parámetros (3PL)[2] asume que la CCI se define mediante la expresión:

$$CCI_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-1,7a_i(\theta - b_i)}} \qquad (2)$$

donde c_i es el *factor de adivinanza*, b_i es la *dificultad* del ítem y a_i es el *factor de discriminación*. El *factor de adivinanza* es la probabilidad de que un alumno sin conocimientos conteste al ítem correctamente (seleccionando la respuesta al azar). La *dificultad* del ítem representa el nivel de conocimiento para el cual la probabilidad de que el alumno falle es igual a la de que acierte más el factor de adivinanza. El *factor de discriminación* es proporcional a la pendiente de la curva. Cuanto mayor es su valor mejor será capaz el ítem de discernir entre los alumnos con mayores niveles de conocimiento de los que tienen menores niveles.

Basándonos en los TAI y la TRI, en nuestro grupo se ha desarrollado la herramienta de tests adaptativos, a través de la Web, SIETTE. A diferencia de la TRI tradicional, el nivel de conocimiento en SIETTE es una variable que puede tomar $n + 1$ valores $v_0 < ... < v_n$ (que en lo sucesivo representaremos por $0, 1, ..., n$). De este modo, en SIETTE, las CCI son vectores de $n + 1$ elementos y la aplicación de la regla de Bayes se reduce a un producto de $n + 1$ valores seguido de un proceso de normalización.

3. Introducción de pistas en un modelo basado en TAI

Para introducir las pistas en nuestro modelo, definamos en primer lugar algunos términos:

- *Ítem.* Utilizamos este término para representar de modo genérico una pregunta o ejercicio que el alumno debe resolver. La solución de esta pregunta o tarea puede darse de diversas maneras: seleccionando una o más opciones de entre un conjunto de respuestas posibles dado un enunciado, escribiendo directamente una breve respuesta, etc.
- Un *test* es una sucesión de ítems.
- *Pista.* Una ayuda o pista es una información adicional que se presenta al alumno

tras proponerle un ítem y antes de que lo conteste. Puede ser una explicación más detallada del enunciado, una ayuda que le permita descartar una o más respuestas, una indicación de cómo proseguir, etc. Las pistas se invocan de modo *activo* (el estudiante puede pedirlas sin más que pulsar un botón) o *pasivo* (el sistema selecciona y presenta la pista de acuerdo al comportamiento del alumno, cuando detecta que éste tiene dificultades, por ejemplo, si lleva mucho tiempo esperando la respuesta del alumno).

Veamos un ejemplo: Consideremos el siguiente ítem:

¿Cuál es el resultado de la expresión: $1/8 + 1/4$*?*
a) $3/4$ *b)*$2/4$ *c)*$3/8$ *d)*$2/8$

Algunas pistas posibles para este ítem podrían ser:

Pista 1. $1/4$ *es equivalente a* $2/8$.
Pista 2. *Primero encuentra fracciones equivalentes con el mismo denominador.*

Una hipótesis simplificadora que ha sido necesario realizar en esta primera etapa de nuestra investigación es que las pistas no modifican el conocimiento del alumno (es decir, que el conocimiento del alumno permanece constante durante el test, incluidas las pistas). Ésta es una hipótesis controvertida pero habitual en los TAI, ya que hace que el modelo sea computacionalmente tratable. En nuestro caso, esta hipótesis puede interpretarse como que las pistas no cambian el conocimiento del alumno, sino que modifican la forma de la curva característica del ítem al que están asociadas. En este sentido, la pista lo que hace es acercar al ítem que actualmente está en la ZDP, hasta el nivel de conocimiento del alumno. Así, cabe considerar que la combinación *ítem+pista* podría considerarse como un nuevo ítem (virtual), que puede medirse y tratarse del mismo modo que los otros ítems en el test; es decir, el nuevo ítem tiene asociada una nueva curva característica cuyos parámetros pueden ser estimados utilizando las técnicas tradicionales de calibración. Sin embargo, ambas curvas no son independientes, sino que se deben satisfacer una serie de relaciones entre ellas. En primer lugar, la pista debe hacer que el ítem sea más fácil. Esta condición se puede establecer en términos matemáticos mediante el siguiente axioma:

Axioma 1. Dado un ítem q y una pista h, para todos los niveles de conocimiento k se debe satisfacer que $CCI_q(k) \leq CCI_{q+\{h\}}(k)$, donde CCI_q representa la curva original y $CCI_{q+\{h\}}$ representa la curva característica de la combinación *ítem+pista*.

Asimismo, si el examinando hace uso de una combinación de pistas, la cuestión debería ser aún más fácil. Matemáticamente esta condición puede expresarse así:

Axioma 2. Dado un ítem q, un conjunto de pistas H y una pista $h \notin H$, para todos los niveles de conocimiento k se debe satisfacer que $CCI_{q+H}(k) \leq CCI_{q+H+\{h\}}(k)$.

Para un conjunto de ítems y pistas, si tras la calibración de los parámetros de las CCI de los ítems virtuales y reales, las curvas resultantes no satisfacen los dos axiomas anteriores, significa que alguna de las informaciones suministradas como pistas no es tal, sino más bien un elemento que confunde al alumno en lugar de ayudarlo, y por tanto debe ser descartado. Este enfoque es simple, pero proporciona un método empírico útil que permite la validación de las pistas propuestas.

En entornos adaptativos, tiene sentido buscar criterios que permitan la selección de la mejor pista disponible. Bajo el marco de la ZDP, si el alumno no es capaz de contestar correctamente al ítem pero éste se sitúa en su ZDP, la mejor pista disponible sería aquélla que trae al ítem I desde la ZDP a la zona del conocimiento del alumno, y por supuesto dependerá de cómo de "lejos" esté la pista en la ZDP; es decir, cuanto más lejos esté la pista del alcance del alumno, según su conocimiento, más detallada, completa o explicativa deberá ser la pista. De este modo, si por ejemplo un ítem I tiene asociada tres pistas h_1, h_2 y h_3 a diferentes niveles de detalle, significa que cada una de ellas es apropiada para una zona diferente de la ZDP, tal como se representa en la Figura 1.

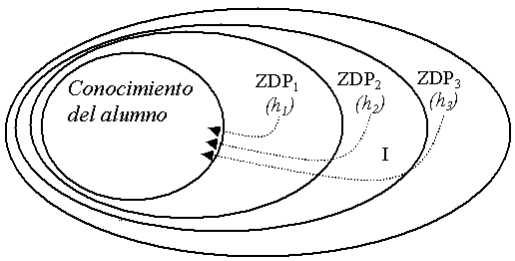

Figura 1. Conocimiento del alumno, ZDP y pistas.

La selección de h_i como la mejor pista a presentar, significaría que el ítem I está en la zona ZDP_i para este alumno. En el ejemplo de la Figura 1, la mejor pista es h_2 ya que el ítem está en la zona ZDP_2. Una posibilidad de selección adaptativa de pistas es usar los mecanismos de selección adaptación de ítems empleados en los TAI. Por consiguiente, dada la estimación de conocimiento de un alumno $\theta(k)$, dadas dos pistas h_1, h_2, y las respectivas curvas características $CCI_{q+\{h1\}}(k)$ y $CCI_{q+\{h2\}}(k)$, la mejor pista es aquélla que minimiza la varianza esperada de la distribución de probabilidad a posteriori. Este mecanismo es simple de implementar y no requiere modificaciones sustanciales en el procedimiento adaptativo, porque el test se usa como mecanismo de evaluación y no como herramienta de aprendizaje. Sin embargo, el uso de mecanismos adaptativos de selección de pistas en este contexto puede servir de estímulo para el alumno, y por tanto mejorar su grado de confianza en sí mismo.

4. Experimentos realizados con alumnos reales

Un paso importante para la integración de este nuevo enfoque de administración adaptativa de pistas en el sistema SIETTE es la calibración de las curvas características asociadas a cada combinación *ítem+pista*. Éste es un objetivo difícil, y para ello se debe seguir una metodología que consta de los siguientes pasos:

1. En primer lugar, hay que desarrollar un banco de ítems y varias pistas para cada uno de ellos.
2. En segundo lugar, hay que presentar los ítems al alumno mediante un test no adaptativo, y calibrar las CCI de los ítems sin pista y las correspondientes a los ítems virtuales, es decir a los pares *ítem+pista*.
3. Finalmente, una vez inferidas las CCI, puede realizarse la administración adaptativa del test y de las pistas asociadas a sus ítems.

En cuanto al primer paso, se ha desarrollado un banco de ítems para la asignatura *Procesadores de Lenguajes* impartida en la E.T.S. de Ingenieros en Informática de la Universidad de Málaga. Cada uno de los ítems tiene entre dos y cuatro pistas asociadas. A continuación se muestran dos de los ítems (junto con las pistas asociadas a ellos) incluidos en el banco de la asignatura:

1. *Indicar la salida que produce el siguiente programa LEX con la entrada* $|*abc*|$
ab/c { printf("uno"); }
c { printf"dos"); }
abc { printf("tres"); }

a) tres b) uno dos c) uno d) uno dos tres

Pista 1. *El contenido de yytext no incluye los caracteres correspondientes a la parte de la expresión que hay después del operador de lectura avanzada "/".*
Pista 2. *Cuando hay un operador de lectura avanzada "/", la longitud del lexema que se forma es la correspondiente a la expresión situada a la izquierda del operador.*

2. *Sea T el conjunto de todos los caracteres ASCII del 1 al 127. Para expresar en LEX el conjunto de todas las cadenas que pueden formarse con los símbolos de T, puede usarse la expresión:*

$a)(.|)*$ $b)[a-zA-Z0-9]*$ $c).*$ $d)[.]*$

Pista 1. *El operador .(punto) representa a cualquier carácter ASCII, exceptuando el fin de línea.*
Pista 2. *El operador .(punto) no tiene ningún significado.*
Pista 3. *El alfabeto ASCII incluye no solo letras y números, sino también caracteres especiales, operadores, signos de puntuación, etc.*

En relación con el segundo paso, se han realizado tres experimentos sobre un total de 263 alumnos matriculados en la asignatura, durante los cursos 2003/2004, 2004/2005 y 2005/2006 (el tamaño de los grupos fue de 100, 80 y 83 alumnos, respectivamente). Todos los alumnos fueron calificados mediante un examen tipo test no adaptativo consistente en 20 ítems, administrado a través de SIETTE. La duración total de la sesión de evaluación estaba limitada a un máximo de 45 minutos. La mayoría de los alumnos, un 87 %, completaron el test, y un 97 % contestó al menos a 18 ítems. Todos los datos se han usado para estos análisis. El test ofrecía la posibilidad de solicitar pistas para cada uno de los ítems. Se trataba de un test heterogéneo en el que a los alumnos se les administraron diversos tipos de ítems (de opción múltiple con respuesta única, de opción múltiple

con respuesta múltiple y de respuesta libre). Los mismos 20 ítems se presentaban en permutaciones diferentes a distintos alumnos para dificultar la copia entre ellos. Una vez que un alumno solicitaba una pista, ésta era seleccionada aleatoriamente dentro del conjunto de pistas disponibles para ese ítem.

Antes de comenzar el test se informó a los alumnos del mecanismo seguido para la puntuación de los ítems. En los casos de ítems de opción múltiple con respuesta simple, se sumaba 1 punto si la respuesta era acertada, y se restaban $1/(n-1)$ (siendo n el número de opciones) puntos, en caso de fallo. Para los ítems de opción múltiple con respuesta múltiple, la puntuación era la suma de $1/n$ puntos positivos por cada opción correcta y los mismos puntos negativos, en caso de opción incorrecta. Para los ítems de respuesta libre, se sumaba 1 punto en caso de que el alumno escribiera la opción correcta y no se penalizaba el fallo. En todos los tipos de ítems no se penalizaba la respuesta en blanco.

Igualmente se informó a los alumnos que el uso de una pista suponía una penalización en la valoración que recibía la respuesta correcta (sobre el máximo de 1 punto por ítem). En el primer experimento, el alumno obtenía un máximo de $0,5$ puntos por respuesta correcta en la que hubiese utilizado una ayuda, mientras que en el segundo obtenía $0,75$, y por último, en el tercero se obtenía 1 punto (es decir, no había penalización por el uso de pistas).

Por ejemplo, si un alumno del segundo experimento responde correctamente 4 de las 5 opciones de un ítem de opción múltiple, tras recibir una ayuda, obtendría en ese ítem una puntuación de $0{,}75 * (4/5 - 1/5) = 0{,}45$, en vez de los 0,60 puntos que habría obtenido si hubiera acertado sin solicitar la ayuda.

Esta evaluación heurística se ha usado como paso inicial para el proceso iterativo de calibración [6]. Como resultado de este proceso se obtienen, además de las curvas características de los ítems y de los pares *ítem+pista*, una nueva calificación del alumno. Esta nueva evaluación basada en la TRI ofrece un mecanismo de evaluación fiable, independiente de la población, y de las puntuaciones heurísticas iniciales[1]. No obstante, los datos anteriores son relevantes desde el punto de vista subjetivo a la hora de decidir si se solicitan pistas o no durante la realización del test.

4.1. Impresión subjetiva del uso de pistas en el primer experimento

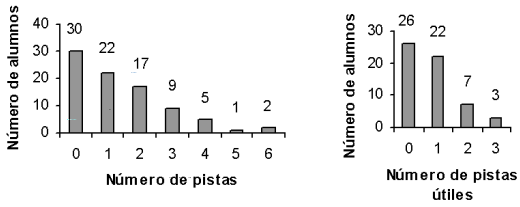

Figura 2. Resultados de los cuestionarios de los alumnos del curso 2003/04.

Durante el transcurso del experimento con el primer grupo (curso 2003/2004), una vez que los alumnos habían acabado el examen, rellenaban un cuestionario sencillo en el cual informaban de cuántas pistas habían solicitado y cuántas de ellas les habían resultado útiles para seleccionar la respuesta correcta. Los resultados se presentan en la Figura 2. En ésta se observa un uso escaso de las pistas, que se explica por el alto factor de penalización. Podemos ver también que los alumnos no consideraron que las pistas fuesen demasiado útiles (según los datos de la encuesta, tan sólo un 65 % de las pistas eran útiles), probablemente debido a que la selección de pistas era aleatoria. Esta administración aleatoria de pistas es precisamente lo que intentamos mejorar con nuestra propuesta.

4.2. Resultados objetivos del uso de pistas

En la Figura 3 se muestran los porcentajes medios totales de pistas solicitadas en cada ítem, por cada uno de los experimentos realizados en los tres cursos académicos. El uso de pistas totales fue del 8, 7 y 53 %, respectivamente, lo que indica que los alumnos sólo perciben de forma cualitativa la penalización que se aplicaba por el uso de pistas, ya que no hubo diferencias significativas entre los

[1]El proceso de calibración consiste en calcular las curvas características a partir de una puntuación inicial, y a partir de ellas una nueva puntuación para el alumno. El proceso itera hasta que se alcanza un equilibrio. Por tanto, los alumnos que han usado pistas (o dicho de otra forma, presumiblemente han contestado a ítems más "fáciles"), obtienen una calificación similar a la que habrían obtenido de no usarlas, ya que las curvas características tienen en cuenta la mayor o menor "dificultad" de los ítems.

dos primeros cursos, a pesar de que la penalización por el uso de pistas en el segundo año era menor. El uso de pistas sólo se generalizó en el tercer experimento, en el que las éstas no penalizaban.

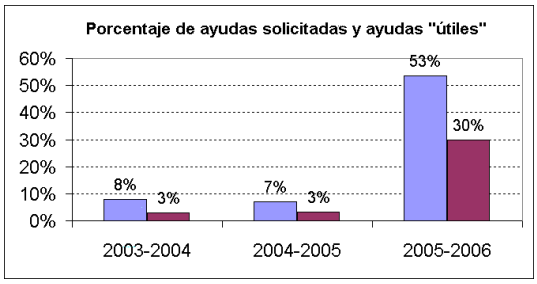

Figura 3. Resultados objetivos del uso de pistas en los cursos de los experimentos.

En cuanto a la utilidad real de las pistas, entendiendo este concepto como el porcentaje de ítems acertados tras solicitar una pista, se ha mantenido en torno a la mitad (38 %, 47 % y 56 %, respectivamente) de las pistas utilizadas. Obsérvese que este dato es incluso más bajo que la percepción subjetiva puesta de manifiesto por los propios alumnos del primer experimento, que fue del 65 %. Esto se explica por el hecho de que, en este caso, las condiciones para considerar "útil" una pista son aún mas restrictivas, ya que implican resolver completamente la cuestión, y no sólo pensar que se ha hecho así, sin tener en cuenta respuestas parcialmente correctas (por ejemplo, en el caso de ítems de respuesta múltiple).

En el conjunto del experimento, la media de respuestas acertadas fue del 55 %, sin utilizar pistas, y del 52 %, en los ítems en los que se utilizaron. Este dato no es muy significativo en sí, pero evidentemente, para los fines del test y del propio experimento, es conveniente garantizar que se mueve en valores medios, por varias razones. En principio, podríamos asumir que el alumno que solicita una pista lo hace porque el ítem le resulta más difícil, o no está completamente seguro de la respuesta correcta (aún en el caso en que no había penalización por las pistas, la lectura de las mismas conlleva un tiempo que estaba limitado en el test). Por otro lado, si las pistas son muy explícitas, el porcentaje de aciertos sería muy alto, y al contrario. Es interesante estudiar no sólo los valores medios de utilidad de las pistas, sino su distribución. Estos datos han sido representados en la Figura 4. El eje de abscisas representa el porcentaje de utilidad de las pistas, y el eje de ordenadas el número de pistas correspondiente. Así, por ejemplo, según la figura, 17 pistas fueron útiles en el intervalo comprendido entre el 60 y el 70 % de los casos. A partir de la resultados expresados en la figura, se observa que la utilidad de las pistas es variable, aunque la mayoría ocupa valores centrales, lo que da entrada al siguiente estudio.

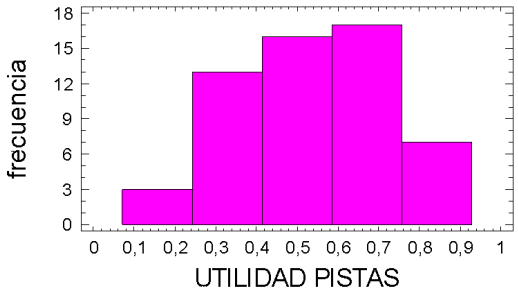

Figura 4. Distribución de la utilidad de las pistas.

4.3. Estudio de la influencia de las ayudas

De los 20 ítems que componen el banco de preguntas, hemos seleccionado dos para mostrar los resultados de este estudio. Ambos corresponden a los ítems que se mostraron al comienzo de esta sección. Para cada uno de ellos, estudiamos la influencia de las pistas mediante la técnica de obtención de curvas características. Tal y como se ha mencionado anteriormente, SIETTE utiliza una aproximación discreta de la TRI mediante N niveles de conocimiento. En este caso inicial, hemos considerado solamente dos niveles: El nivel 1 está formado por aquellos alumnos cuya puntuación en el test fue menor de 50 puntos (sobre un máximo de 100 puntos), y el nivel 2, por los restantes (puntuación superior a 50 puntos).

Con este objetivo, los alumnos de los tres experimentos se agruparon según su rendimiento en el test. En las tablas 1 y 2 vemos los datos correspondientes a los dos ítems de ejemplo que se han presentado: el ítem 1, con dos pistas asociadas, y el ítem 2, con tres pistas. Las tablas muestran los resultados por cada ítem, es decir, el número de alumnos que resolvieron correctamente el ítem en cada uno de los niveles. Así por ejemplo, el 25/83 de la primera posición de la primera fila significa que, de entre aquellos alumnos clasificados en el nivel 1, 83 contestaron al ítem 1 sin pistas y 25 de ellos fueron capaces de resolverlo correctamente; 7/13 significa que a 13 de los estudiantes

que pidieron pistas se les presentó la pista 1, y 7 de ellos contestaron correctamente; y así sucesivamente. Estos mismos datos se representan en porcentajes en la Figura 5.

Ítem 1	*Sin pista*	*Pista 1*	*Pista 2*	
Nivel 1	25/83	7/13	5/20	
Nivel 2	83/115	13/16	20/23	
Ítem 2	*Sin pista*	*Pista 1*	*Pista 2*	*Pista 3*
Nivel 1	43/74	8/16	2/8	5/7
Nivel 2	91/102	12/15	11/18	15/20

Tabla 1. Proporción de alumnos que respondieron de forma correcta.

Ítem 1	*Sin pista*	*Pista 1*	*Pista 2*	
Nivel 1	30,1 %	53,8 %	45,5 %	
Nivel 2	72,2 %	81,3 %	87 %	
Ítem 2	*Sin pista*	*Pista 1*	*Pista 2*	*Pista 3*
Nivel 1	58,1 %	50 %	25 %	71,4 %
Nivel 2	89,2 %	80 %	61,1 %	75 %

Tabla 2. Porcentaje de alumnos que respondieron de forma correcta.

Somos conscientes de que, desafortunadamente, el uso de la penalización ha hecho que el número de casos de pistas utilizadas en algunos casos sea demasiado bajo (véanse los intervalos de confianza en la siguiente subsección) para permitirnos establecer conclusiones que sean estadísticamente significativas, pero pensamos que nos pueden servir como una primera aproximación para determinar el tipo de información que puede obtenerse mediante el uso de estos análisis. Así, por ejemplo, en la Figura 5 se observa que, para el ítem 1, las dos pistas son más o menos igual de útiles y ayudan tanto a los alumnos clasificados en el nivel 1 como a los clasificados en el nivel 2. Para el ítem 2, la pistas 1 y 2 resultan inútiles, e incluso confunden a los alumnos. Por el contrario, la pista 3 parece que ayuda a los alumnos de nivel más bajo pero confunde a los más avanzados. Esto sugiere que este tipo de análisis permitiría eliminar las pistas "malas" (las que confunden a ambos grupos, las que no tienen efectos positivos, y aquéllas que tienen un efecto demasiado positivo) del banco de pistas, mejorando así su calidad. Del mismo modo, las pistas pueden seleccionarse de modo adaptativo, dependiendo de la estimación actual del nivel de conocimiento del alumno para satisfacer mejor sus necesidades.

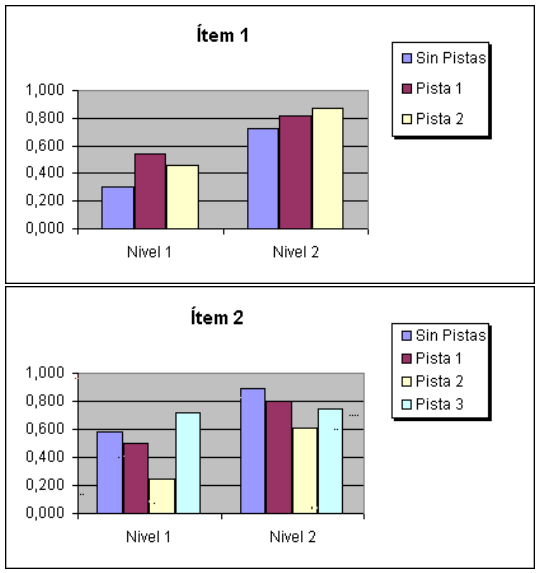

Figura 5. Porcentajes de respuestas correctas en función del nivel de conocimiento de los alumnos.

Estos mismos datos pueden utilizarse para hacer aproximaciones a las CCI, tal y como se muestra en la Figura 6. En ella podemos ver claramente que todas las pistas del ítem 1 cumplen el *Axioma 1*, que establece que para cualquier nivel de conocimiento k se debe satisfacer que $CCI_q(k) \leq CCI_{q+\{h_k\}}(k)$. Sin embargo, esto no ocurre en el ítem 2. Por tanto, si tal hipótesis se considera razonable, las pistas que no la cumplen deberían ser descartadas.

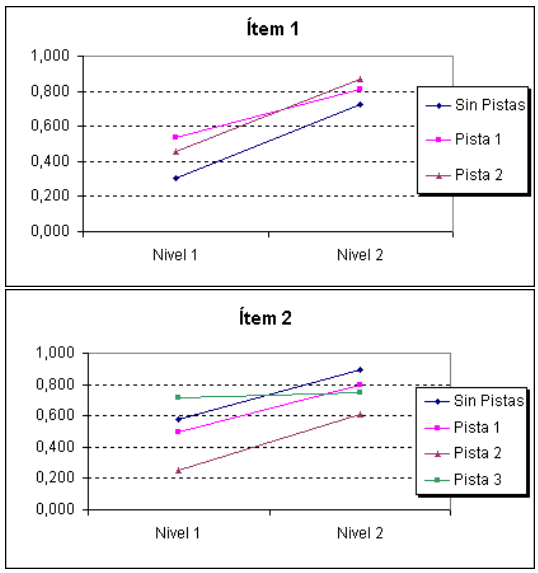

Figura 6. Aproximaciones de las CCI.

4.4. Intervalos de confianza

Hemos analizado los intervalos de confianza de los porcentajes obtenidos mediante la fórmula aproximada del *error esperado* (EE) para poblaciones pequeñas, esto es:

$$EE = z\sqrt{\frac{P(1-P)}{n}} + \frac{0,5}{n} \qquad (3)$$

En donde n es el tamaño de la muestra, P es la proporción (media) de la muestra y z se obtiene de la distribución normal, expresada en la Tabla 3.

Ítem 1	*Sin pista*	*Pista 1*	*Pista 2*
Nivel 1	0,105	0,309	0,340
Nivel 2	0,086	0,223	0,159

Ítem 2	*Sin pista*	*Pista 1*	*Pista 2*	*Pista 3*
Nivel 1	0,119	0,276	0,363	0,406
Nivel 2	0,065	0,236	0,253	0,215

Tabla 3. Amplitud de los intervalos de confianza de las proporciones.

La Tabla 3 muestra la amplitud de los intervalos de confianza habituales al 95 % (z=1, 96). Como puede apreciarse, en la mayoría de los casos los intervalos son demasiado grandes como para establecer conclusiones estadísticamente significativas, a este nivel de confianza.

4.5. Análisis basado en la calibración de los ítems según el modelo 3PL

Mucho más interesante resulta realizar un análisis completo de las curvas características mediante la aplicación de la TRI. Para ello, supondremos que las CCI siguen el modelo logístico de tres parámetros, 3PL, expresado en la ecuación 2.

Para el experimento, se ha llevado a cabo la calibración de las curvas características de 81 ítems: las de los 20 originales, y además, las de los 61 virtuales que se obtienen mediante la conjunción *ítem+pista*. Con este fin, se ha utilizado el programa MULTILOG [13]. Los resultados han evidenciado que la distribución de los niveles de conocimiento de los alumnos (θ) resulta ser normal con media $0,018$ y desviación típica $0,337$; es decir, el $99,7\,\%$ de la muestra tiene valores de θ en el intervalo $[-1, 1]$. En la Figura 7, se muestran los resultados para los ítems 1 y 2, usando este intervalo. En el eje de abscisas se muestra el nivel de conocimiento (θ), y en el eje de ordenadas la probabilidad de acertar el ítem. Las curvas de los ítems sin pistas se muestran mediante una línea de puntos.

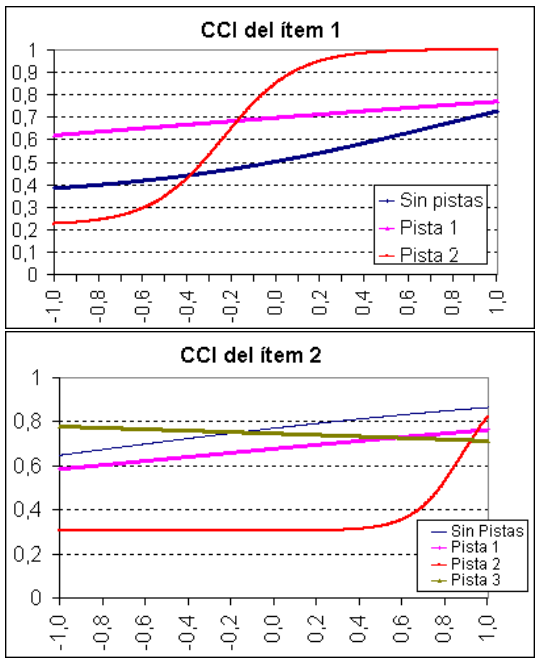

Figura 7. CCI de los ítems reales y virtuales.

Como se puede apreciar, tras este análisis, las pistas del ítem 1 resultan útiles sólo a partir de un determinado umbral de conocimiento, el cual se obtiene mediante la intersección de ambas curvas. La pista 1 resulta de utilidad para casi todos los alumnos, mientras que la pista 2, es útil a partir de los niveles de conocimiento que cumplen que $\theta > -0,35$ (punto de corte con la curva del ítem sin pista); es decir, para el $84\,\%$ de los alumnos. En cuanto al ítem 2, puede observarse que la pista 1 no supone ninguna mejora en el porcentaje de aciertos, y su curva es muy similar a la del ítem original, por lo que puede deducirse que contiene una información irrelevante para resolver el ítem. La pista 2, perjudica el resultado casi todos los alumnos. Por último, la pista 3 es irrelevante, o incluso puede resultar confusa, como se ve claramente por su pendiente negativa y, por esta razón, debería ser eliminada.

Desgraciadamente, como ya se ha visto en el estudio anterior, los datos recogidos no son suficientes

para garantizar un buen ajuste de las curvas características, aunque los resultados no afecta a las conclusiones que se proponen en este artículo.

Estos resultados coinciden con los que se habían obtenido anteriormente al aplicar sólo dos niveles. Si bien son más detallados, no por ello son más precisos y están sujetos a la incertidumbre estadística en la calibración de las curvas. Por otra parte, el análisis mediante la TRI clásica, presupone que las curvas características siguen el modelo 3PL, lo cual no es necesariamente cierto.

5. Conclusiones y trabajo futuro

En este artículo se han presentado algunas ideas sobre la selección adaptativa de pistas en un entorno de evaluación basado en TAI. Como se ha mencionado anteriormente, este tipo de tests se basan en la TRI.

Es necesario hacer especial hincapié en que, en esta aproximación, las pistas no se consideran modificadores del conocimiento del alumno, sino más bien modificadores de las CCI. Algunos axiomas formales que todo modelo de pistas deben satisfacer, han sido presentados y justificados de un modo informal. Un estudio preliminar ha sugerido que el uso de las pistas, en estos entornos, es adecuado y factible. Asimismo, el estudio anterior ha sido validado mediante la calibración de las CCI para cada par *ítem+pista*, utilizando datos empíricos. Las curvas obtenidas permiten discernir entre aquellas pistas que resultan útiles y las que no, y permite determinar qué pistas deben ser eliminadas por resultar confusas a los alumnos. Desgraciadamente, los datos obtenidos no son, por el momento, lo suficientemente concluyentes, debido a los posibles errores estadísticos que se derivan del tamaño de las muestras poblacionales utilizadas en los experimentos.

Una vez se obtengan resultados más fiables, éstos servirán de base para la integración e implementación de este modelo en SIETTE, y para permitir la selección de adaptativa tanto de ítems como de pistas en nuestro sistema de tests.

Referencias

[1] I. Arroyo, J. E. Beck, B. P. Woolf, C. R. Beal, and K. Schultz. Macroadapting animalwatch to gender and cognitive differences with respect to hint interactivity and symbolism. In *Proceedings of the 5th World Conference of Intelligent Tutoring Systems. ITS'00*, pages 604–614. Springer-Verlag, 2000.

[2] A. Birnbaum. Some latent trait models and their use in inferring an examinee's mental ability. In *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley, 1968.

[3] S. Chua Abdullah. *Student Modelling by Adaptive Testing - A Knowledge-based Approach*. PhD thesis, University of Kent, Canterbury, June 2003.

[4] R. Conejo, E. Guzmán, E. Millán, M. Trella, J. L. Pérez de la Cruz, and A. Ríos. Siette: a web-based tool for adaptive testing. *Journal of Artificial Intelligence in Education*, 14:29–61, 2004.

[5] R. Conejo, E. Guzmán, J. L. Pérez de la Cruz, and E. Millán. Introducing adaptive assistance in adaptive testing. In C. K. Looi, G. McCalla, B. Bredeweg, and J. Breuker, editors, *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology. (AIED 2005)*, pages 777–779. Amsterdam: IOS Press, 2005.

[6] S. E. Embretson and S. P. Reise. *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum, 2000.

[7] A. S. Gertner, C. Conati, and K. VanLehn. Procedural help in andes: Generating hints using a bayesian network student model. In *Proceedings of the 15th National Conference on Artificial Intelligence*. Madison, Wisconsin, 1998.

[8] E. Guzmán and R. Conejo. A brief introduction to the new architecture of siette. In P. De Bra and W.Ñejdl, editors, *Proceedings of the IIIth International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems(AH 2004). Lecture Notes in Computer Science*, number 3137, pages 405–408. New York: Springer Verlag, 2004.

[9] R. K. Hambleton, J. Swaminathan, and H. J. Rogers. *Fundamentals of Item Response Theory*. Sage publications, 1991.

[10] G. D. Hume, J. Michael, A. Rovick, and M. W. Evens. Hinting as a tactic in one-on-one tutoring. *Journal of Learning Sciences*, 5(1):23–47, 1996.

[11] E. Millán, , E. García-Hervás, E. Guzmán, A. Rueda, and J. L. Pérez de la Cruz. Tapli: An adaptive web-based learning environment for linear programming. In R. Conejo, M. Urretavizcaya, and J. L. Pérez de-la Cruz, editors, *Current Topics in Artificial Intelligence. Lecture Notes in Arficial Intelligence*, number 3040, pages 676–687. New York: Springer-Verlag, 2004.

[12] L. M. Rudner. An examination of decision-theory adaptive testing procedures. In *Annual meeting of the American Educational Research Association*, 2002.

[13] D. Thissen. Multilog: Multiple, categorical item analysis and test scoring using item response theory (versión 5.1), 1988.

[14] L. Vygotskii. *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press, 1978.

[15] H. Wainer. *Computerized Adaptive Testing: A Primer*. Lawrence Erlbaum, Hillsdale, NJ, 1990.

[16] G. Wells. *Dialogic Inquiry: Towards a Socio-Cultural Practice and Theory of Education*. New York: Cambridge University Press, 1999.