Un Estudio sobre la dificultad de los ítems en tests de Informática

R. Conejo, E. Guzmán de los Riscos, J. L. Pérez de la Cruz

Lenguajes y Ciencias Computación

Universidad de Málaga

ETSI Inf, Bulevar Luis Pasteur

29071 Málaga

{conejo,guzman,perez@lcc.uma.es}

Resumen

Se presenta un estudio empírico de la exactitud de profesores y alumnos en la tarea de estimar la dificultad de preguntas tipo test en dos asignaturas de contenido informático de la Universidad de Málaga. Se calcula la coherencia de cada uno de los dos grupos por separado y en conjunto, así como diferentes medidas de la correlación entre las diversas estimaciones y los datos empíricamente obtenidos.

1. Introducción

El nuevo modelo de enseñanza y aprendizaje que se pretende implantar en la universidad fomenta el empleo de técnicas de evaluación continua de los conocimientos y competencias del estudiante. Quizás por ello, crece el número de asignaturas en las que la administración de tests es parte más o menos importante del proceso de evaluación.

Sin embargo, para que un test sea una herramienta útil se requieren cierta habilidad y esfuerzo por parte del profesor (vd. por ejemplo [1]); y, como no es seguro que se den estas condiciones, se ha llegado a recomendar a los profesores noveles: *"No hagas exámenes de respuesta múltiple"* [3].

En el presente trabajo no pretendemos abogar por el empleo o el rechazo de los tests en la evaluación de asignaturas de Ingeniería Informática. Nuestro objetivo, más modesto, es analizar de forma empírica si los docentes (y los discentes) son buenos jueces del grado de dificultad que presentan los ítems de un test (se denomina "ítem" a cada una de las preguntas de cualquier tipo que constituyen un test). Una respuesta negativa a esta cuestión, como parecen sugerir los datos, será solamente un elemento más a tener en cuenta para extremar la prudencia en la administración y análisis de este tipo de pruebas.

En la literatura psicométrica existen varios trabajos que estudian hasta qué punto los "expertos" en una materia estiman adecuadamente la dificultad de un ítem. Por ejemplo, en [11] se propuso a 59 profesores de matemáticas de los grados 7 al 11 (alumnos aprox. de 13 a 17 años) que juzgaran el porcentaje de sus alumnos que responderían correctamente a cada una de las preguntas de un conjunto de 50 tomadas de tests preexistentes. Posteriormente se administraron las preguntas a sus alumnos y se calcularon los porcentajes reales, así como los coeficientes de correlación entre los porcentajes estimados y los reales. La proporción de profesores para los cuales –al nivel de significación del 0,05– no podía afirmarse que esta correlación fuera distinta de 0 varió entre el 40 % y el 66 %.

En [2], [6] se describen dos experimentos llevados a cabo en el Educational Testing Service (ETS). Los ítems provenían de un test

para comprobar el dominio del inglés escrito (TSWE, un test complementario del test SAT que se administra a los candidatos al ingreso en la Universidad) [2] y de la parte general (ítems de analogías) del GRE (test que se administra a los candidatos a la admisión en másteres universitarios). Todos los ítems se extrajeron de tests ya revelados, y los porcentajes de aciertos en la población objetivo habían sido calculados. Se pidió a varios expertos del ETS que estimaran el porcentaje de aciertos de cada pregunta y se calcularon las correlaciones para cada experto y cada tipo de pregunta. Los datos sugerían que la exactitud de la estimación de los expertos no se aproximaba a la que sería necesaria para prescindir de la calibración empírica del grado de dificultad de cada ítem.

Por último, en [10] se seleccionaron aleatoriamente 26 profesores de Ciencias Naturales de 6^o grado y se les propuso que estimaran el porcentaje de sus alumnos que responderían correctamente a cada uno de los 50 ítems de un test de su materia, calculándose posteriormente los porcentajes reales obtenidos al administrar el test. Las diferencias para cada profesor variaron entre 15,4 % por defecto (estimación del profesor inferior al porcentaje real) y 24,2 % por exceso (estimación del profesor superior al porcentaje real). Comparando sus hallazgos con los de otros investigadores, la conclusión de [10] es que "nuestros resultados fueron consistentes con las investigaciones anteriores que sugerían que estimar con exactitud la dificultad de un ítem es bastante difícil".

En la línea de los trabajos citados, en el que aquí se describe hemos investigado si los profesores de dos asignaturas de la Universidad de Málaga relacionadas con la programación son buenos jueces de la dificultad de las preguntas efectivamente administradas en sus clases. También hemos evaluado la exactitud de los estudiantes en esta tarea de estimación. En las secciones siguientes describiremos con más detalle las experiencias realizadas (sección 2) y los resultados obtenidos (sección 3) para finalizar con un análisis de éstos (sección 4) y con algunas reflexiones de índole práctica (sección 5).

2. Método

Describiremos las experiencias llevadas a cabo en dos asignaturas de programación en la Universidad de Málaga. Ambas experiencias se realizaron mediante la herramienta SIETTE [4], [7], [8]. SIETTE es un sistema basado en la web que los profesores pueden emplear tanto para proporcionar tests de autoevaluación como para administrar tests de calificación. También proporciona facilidades adicionales para analizar los resultados de un test. Desde hace varios años, el sistema se viene empleando de forma habitual en diversas asignaturas impartidas en la Universidad de Málaga y en otras universidades españolas. Para este estudio, se seleccionaron dos materias:

-"Inteligencia Artificial e Ingeniería del Conocimiento" (IAIC), correspondiente al 4^o curso de la titulación "Ingeniero en Informática". Esta experiencia se llevó a cabo en febrero de 2004.

-"Elementos de Programación" (EP), correspondiente al 1^{er} curso de la titulación "Ingeniero Técnico de Telecomunicación". Esta experiencia se llevó a cabo en junio de 2007.

El diseño experimental en ambos casos fue el mismo.

Se partió de un banco de ítems multirrespuesta desarrollado por los profesores que impartían la asignatura (3 en ambos casos). SIETTE soporta ítems de muy diversos tipos (respuesta libre, verdadero/falso, etc.) En nuestro caso se emplearon exclusivamente ítems en los que se ofrecían 3 respuestas posibles, de las cuales una y sólo una era correcta.

Los profesores construyeron de común acuerdo la especificación de un test concreto formado por 20 ítems de este banco.

Antes de administrar el test, y para cada uno de sus ítems, a cada profesor se le planteó la siguiente pregunta: "¿Cuál es el grado de dificultad de este ítem? (Dé un número entero en

[0,10]". La pregunta era voluntariamente ambigua y no aludía a porcentajes de ningún tipo. Sin embargo, ningún profesor solicitó aclaraciones; todos proporcionaron los valores numéricos solicitados.

Posteriormente se administró el test a los estudiantes. En el caso de IAIC, el test era parte de la evaluación de la asignatura. Los alumnos deben superar el test en un entorno controlado (laboratorio de la universidad) en un máximo de 25 minutos. Un total de 43 alumnos se presentaron a realizar el test en febrero de 2004. En el caso de EP, el test se ofreció con carácter voluntario y con fines de autoevaluación, previamente a la realización del examen final de la asignatura. Los alumnos podían realizar el test fuera del laboratorio y repetirlo varias veces (con ciertas limitaciones). Un total de 103 alumnos completaron el test.

La misma pregunta sobre la dificultad de cada ítem se planteó a los estudiantes a los que se había administrado el test, una vez que lo habían completado, pero sin haberles mostrado cuál era la respuesta correcta. Al igual que los profesores, ninguno de ellos solitó aclaraciones sobre el significado del concepto "dificultad". En IAIC proporcionaron estimaciones 14 de los 43 alumnos. En el caso de EP, proporcionaron estimaciones 42 de los 103 alumnos.

A partir de estos datos se estimó el grado de consistencia existente entre los valores de dificultad proporcionados por los diversos sujetos: entre profesores, entre alumnos, y entre profesores y alumnos. Para ello se calculó en cada caso el coeficiente alfa de Cronbach [5]. También se calcularon los coeficientes de correlación de Pearson y sus intervalos de confianza.

Posteriormente se calculó la dificultad real de cada ítem en términos de la teoría clásica del test (CTT) y de diversos modelos de la teoría de respuesta al ítem (RTI)[9]. En este artículo nos referiremos únicamente a la CTT, en la cual la dificultad del ítem i_n se define como la proporción de estudiantes que no ha respondido correctamente a i_n.

Finalmente se compararon las dificultades estimadas por profesores y estudiantes con las dificultades computadas a partir de los datos reales.

3. Resultados

3.1. Test de programación Lisp

3.1.1. Estimaciones de los profesores

Representaremos por P1, P2, P3 a cada uno de los 3 profesores participantes en el experimento, y por mP al promedio de los tres. Las estimaciones de dificultad promediadas sobre las 20 preguntas fueron las siguientes:

P1	P2	P3	mP
5,00	5,50	5,80	5,43

La coherencia entre las estimaciones de los 3 profesores, medida por el coeficiente α de Cronbach, fue $\alpha = 0{,}68$.

Los valores calculados para las correlaciones y sus intervalos de confianza se muestran en el cuadro 1.

	ρ	
P1–P2	0,51	0,09–0,78
P1–P3	0,41	-0,04–0,72
P2–P3	0,34	-0,12–0,68

Tabla 1: Estimaciones de los profesores: correlaciones y sus intervalos de confianza ($\alpha = 0,05$) (Lisp)

A efectos ilustrativos, también se muestra en la figura 1 la nube de puntos para las estimaciones de los profesores P2 y P3.

3.1.2. Estimaciones de los alumnos

Representaremos por mA al promedio de las estimaciones de los alumnos. El valor de mA promediado sobre las 20 preguntas fue 5,66.

La coherencia entre las estimaciones de todos los alumnos, medida por el coeficiente α de Cronbach, fue $\alpha = 0{,}83$.

Considerando las estimaciones promedio mP sobre el conjunto de profesores y mA sobre el conjunto de alumnos, la correlación entre

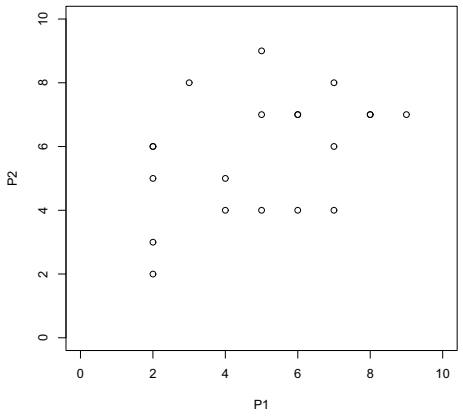

Figura 1: Dificultades estimadas por P1 y P2(Lisp)

ambas es $\rho = 0,52$, y su intervalo de confianza ($\alpha = 0,05$) es $[0,10, 0,78]$.

A efectos ilustrativos, se muestra en la figura 2 la nube de puntos para las estimaciones promedio mP y mA.

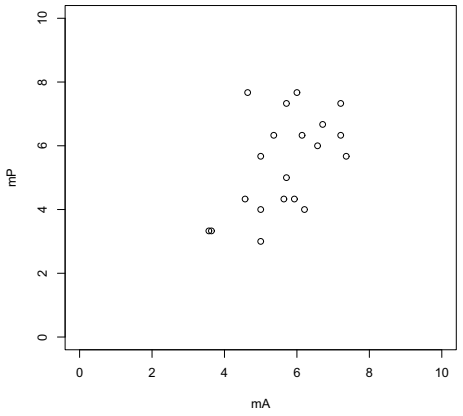

Figura 2: Dificultades promedio profesores y alumnos (Lisp)

3.1.3. Comparación con las dificultades reales

Una vez realizado el test, se estimó su fiabilidad calculando el coeficiente α sobre los vectores de respuesta de todos los alumnos, obteniéndose $\alpha = 0,87$. Comprobada la fiabilidad, para cada ítem se calculó su dificultad real tct, medida como el porcentaje de respuestas correctas proporcionadas sobre el total de alumnos examinados.

Los errores de profesores y alumnos, calculados como la diferencia entre la estimación de dificultad y la dificultad real, aparecen sumarizados en el cuadro 2.

error	P1	P2	P3	mP	mA
medio	1,31	1,81	2,11	1,75	1,97
σ	2,79	2,07	2,65	2,21	1,46
mín.	-2,88	-2,84	-2,84	-2,50	-1,62
máx.	6,21	5,14	5,60	5,47	4,75

Tabla 2: Errores en las estimaciones (Lisp)

Las correlaciones de la dificultad real con las estimaciones promedio de profesores y alumnos aparecen en el cuadro 3, junto con sus intervalos de confianza.

	ρ	
mA–tct	0,64	0,28–0,84
mP–tct	0,18	-0,28–0,58

Tabla 3: Correlaciones dif. estimada/dif. real y sus intervalos de confianza ($\alpha = 0,05$) (Lisp)

A efectos ilustrativos, se muestran en la figura 3 las dos nubes de puntos para las estimaciones promedio mP y mA en función de tct.

3.2. Test de Elementos de Programación

3.2.1. Estimaciones de los profesores

Las estimaciones de dificultad promediadas sobre las 20 preguntas fueron las siguientes:

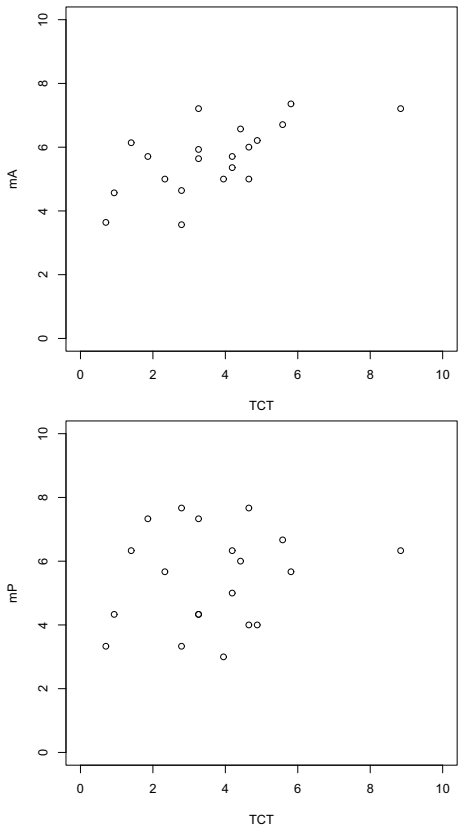

Figura 3: Dificultades estimadas vs. real (Lisp)

P1	P2	P3	mP
6,30	5,35	6,70	6,12

La coherencia entre las estimaciones de los 3 profesores, medida por el coeficiente α de Cronbach, fue $\alpha = 0{,}70$. Los valores calculados para las correlaciones y sus intervalos de confianza se muestran en la tabla 4. Se observa que los resultados son similares a los obtenidos para el test de Lisp.

3.2.2. Estimaciones de los alumnos

El valor de mA promediado sobre las 20 preguntas fue 6,37.

La coherencia entre las estimaciones de todos los alumnos, medida por el coeficiente α de Cronbach, fue $\alpha = 0{,}89$.

	ρ	
P1–P2	0,49	0,06–0,77
P1–P3	0,48	-0,05–0,76
P2–P3	0,33	-0,13–0,64

Tabla 4: Estimaciones de los profesores: correlaciones y sus intervalos de confianza ($\alpha = 0,05$) (Elem. Prog.)

Considerando las estimaciones promedio mP sobre el conjunto de profesores y mA sobre el conjunto de alumnos, la correlación entre ambas es $\rho = -0,24$, y su intervalo de confianza ($\alpha = 0,05$) es $[-0,62, 0,22]$.

A efectos ilustrativos, se muestra en la figura 4 la nube de puntos para las estimaciones promedio mP y mA.

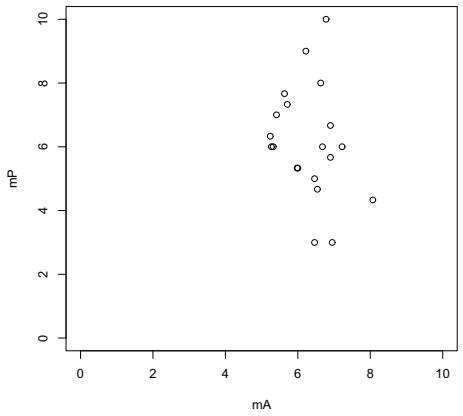

Figura 4: Dificultades promedio profesores y alumnos (Elem. Prog.)

3.2.3. Comparación con las dificultades reales

Una vez realizado el test, se estimó su fiabilidad calculando el coeficiente α sobre los vectores de respuesta de todos los alumnos, obteniéndose $\alpha = 0,89$. Comprobada la fiabi-

lidad, para cada ítem se calculó su dificultad real tct, medida como el porcentaje de respuestas correctas proporcionadas sobre el total de alumnos examinados.

Los errores de profesores y alumnos, calculados como la diferencia entre la estimación de dificultad y la dificultad real, aparecen sumarizados en el cuadro 5.

error	P1	P2	P3	mP	mA
medio	1,14	0,19	1,54	0,96	1,16
σ	2,47	2,20	2,74	2,07	1,61
mín.	-3,52	-5,10	-3,39	-2,10	-2,49
máx.	6,23	3,23	6,23	5,23	3,64

Tabla 5: Errores en las estimaciones (Elem. Prog.)

Las correlaciones de la dificultad real con las estimaciones promedio de profesores y alumnos aparecen en el cuadro 6, junto con sus intervalos de confianza.

	ρ	
mA–tct	0,41	-0,04–0,72
mP–tct	0,31	-0,15–0,66

Tabla 6: Correlaciones dif. estimada/dif. real y sus intervalos de confianza ($\alpha = 0,05$) (Elem. Prog.)

A efectos ilustrativos, se muestran en la figura 5 las dos nubes de puntos para las estimaciones promedio mP y mA en función de tct.

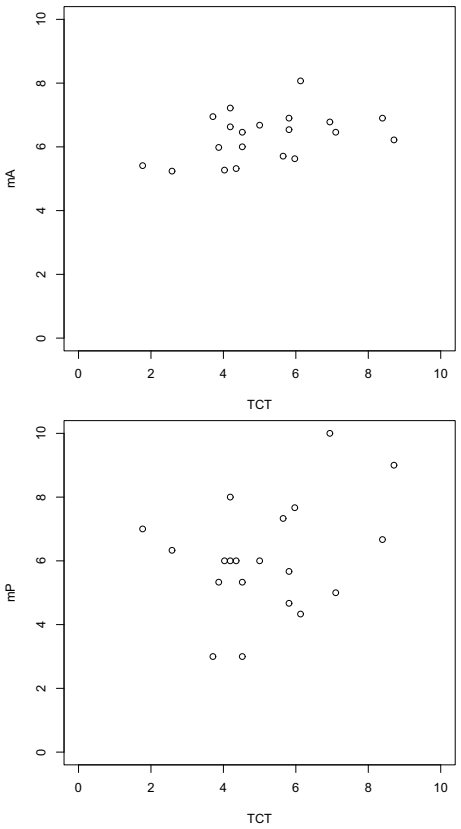

Figura 5: Dificultades estimadas vs. real (Lisp)

4. Discusión

A partir de los datos expuestos, intentaremos responder algunas preguntas sobre las estimaciones de profesores y alumnos.

¿Tienden los alumnos a considerar que la dificultad de un ítem es mayor que la estimada por el profesor? La "sabiduría popular" suele reponder esta cuestión afirmativamente. Sin embargo, los datos recogidos en este estudio sugieren lo contrario. En efecto, las medias para mA y mP no difieren significativamente en ninguna de las dos asignaturas.

Si se consideran varios profesores, ¿estiman éstos de forma uniforme la dificultad de cada ítem?

El estadístico α de Cronbach se suele emplear para medir la coherencia de varias sucesiones de medidas referentes a la misma sucesión de eventos. Convencionalmente se toma $\alpha = 0{,}70$ como el mínimo valor que indica una coherencia aceptable. En el caso de Lisp, $\alpha = 0,68 < 0,70$; en el caso de Elementos, α es exactamente 0,70. Ello parece indicar que

coherencia está en el límite de la aceptabilidad. Intuitivamente, la inspección de los cuadros 1 y 4 y de la figura 1 parecen indicar que la coherencia es pequeña.

Si se consideran varios alumnos, ¿estiman éstos de forma uniforme la dificultad de cada ítem?

La respuesta es afirmativa en ambos casos, ya que los valores son $\alpha = 0,83$ (Lisp) y $\alpha = 0,89$ (Elementos).

¿Están correlacionadas las estimaciones de dificultad que realizan profesores y alumnos?

Para responder a esta pregunta, hemos considerado la relación entre mA y mP. Para el test de Lisp, $\rho = 0,52$ con un intervalo de confianza al 0,05 de [0,10,0,78]. Por tanto, parece que sí se puede afirmar que están relacionadas (figura 2). Para el test de Elementos (figura 4), la respuesta es completamente diferente: $\rho = -0,24$ y el intervalo es [-0,62, 0,22], lo oual significa que en realidad no se ha detectado ninguna correlación entre ambas estimaciones.

¿Son los profesores buenos estimadores de la dificultad de un ítem?

La respuesta es negativa en ambos experimentos. Para el test de Lisp (cuadro 2, figura 3 inferior), la correlación entre mP y la dificultad real es 0,18 (intervalo [-0,28, 0,58]), lo cual significa que en realidad no se ha detectado ninguna correlación entre la dificultad estimada por los profesores y la dificultad real. Algo mayor es el valor para el test de Elementos (figura 5 inferior), pero sigue sin ser significativo.

¿Son los alumnos buenos estimadores de la dificultad de un ítem?

No son muy buenos estimadores, pero en ambos experimentos hay indicios de que son mejores que sus profesores. En el caso del test de Lisp, la correlación entre mA y el valor real (cuadro 2, figura 3 superior) es 0,64 (intervalo [0,28, 0,84]). Para el test de Elementos la correlación calculada entre mA y el valor real (figura 5 superior) es $\rho = 0,41$, mayor que la correspondiente a mP y el valor real, pero no podemos rechazar al 0,05 la hipótesis nula (intervalo de confianza [-0,04, 0,72]). En cualquier caso, Béjar [2] señala que un valor de $\rho = 0,80$ sería el mínimo exigible para aceptar una estimación.

5. Conclusiones

De los resultados expuestos en las secciones anteriores se pueden extraer ciertas conclusiones para la práctica docente.

En primer lugar, podemos decir que los profesores no estiman bien la dificultad de los ítems de un test. Se podría alegar que los resultados de estos dos experimentos no son generalizables, pero hay que considerar que la mayoría de los resultados citados en la literatura están en concordancia con los aquí obtenidos.

En consecuencia, si se pretende administrar un test adaptativo, en el cual las preguntas se seleccionen dinámicamente en función de las respuestas anteriores del alumno, no es posible confiar en las dificultades estimadas por los profesores que crearon el banco de ítems; es inevitable recurrir a la calibración empírica de los mismos, como ya se afirmaba en [2], [6].

Quizás esto no afecte a la mayoría de los docentes, pero hay otro aspecto que sí es relevante para todos: en realidad, ¿qué significa la puntuación obtenida en un test por cada alumno? En ausencia de cualquier estimación fiable sobre la dificultad de cada pregunta, el valor absoluto de esta puntuación carece de significado.

Por tanto, el establecimiento de la nota de corte del test, que convencionalmente se suele establecer en 5/10, es en cierta forma arbitrario.

En efecto, para que esta convención tenga algún significado real, debemos suponer que la nota de corte corresponde al porcentaje de aciertos que va a obtener cierto percentil prefijado de la clase, o al porcentaje de aciertos que obtendría un "alumno mínimamente competente"; porcentajes ambos que dependen de la dificultad real de cada ítem.

Ahora bien, como se ha visto, las dificultades estimadas por diferentes profesores están

débilmente correlacionadas entre sí y con los porcentajes de acierto en la clase, lo cual pone gravemente en duda la verosimilitud de tal suposición. Aún más, promediando sobre las estimaciones de varios profesores tampoco se llega a una estimación satisfactoria.

Referencias

[1] Francisco J. Abad, Jesús Atencia, Carmen García, Pedro Hontangas, Julio Olea, Vicente Ponsoda, Javier Revuelta, Manuel Suero, and Carmen Ximénez. Proyecto de innovación docente: ayuda a la creación de exámenes. http://www.uam.es/docencia/ace/, consultado 2008-02-04.

[2] Isaac I. Bejar. Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7:303–310, 1983.

[3] Agustín Cernuda, Faraón Llorens, Joe Miró, Rosana Satorre, and Miguel Valero. *Guía para el profesor novel (ver. 1.0)*. Editorial Marfil, Alcoy, 2005. ISBN: 84-268-1243-0.

[4] Ricardo Conejo, Eduardo Guzmán, Eva Millán, Monica Trella, J.L. Pérez de-la Cruz, and Antonia Ríos. SIETTE: A web-based tool for adaptive testing. *Journal of Artificial Intelligence in Education*, 14:29–61, 2004.

[5] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334, 1951.

[6] Mary K. Enright and Isaac I. Bejar. An analysis of test writer's expertise: Modeling analogy item difficulty. Technical Report ETS-RR-89-35, Educational Testing Service, Princeton, NJ, July 1989.

[7] Eduardo Guzmán, Ricardo Conejo, and J.L. Pérez de-la Cruz. Adaptive testing for hierarchical student models. *User Modeling and User-Adapted Interaction*, 17:119–157, 2007.

[8] Eduardo Guzmán, Ricardo Conejo, and J.L. Pérez de-la Cruz. Improving student performance using self-assessment tests. *IEEE Intelligent Systems*, 22(4):46–52, 2007.

[9] Ronald K. Hambleton, J. Swaminathan, and H. Jane Rogers. *Fundamentals of Item Response Theory*. Sage Publications, Newbury Park, 1991.

[10] James C. Impara and Barbara S. Plake. Teachers' ability to estimate item difficulty: A test on the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35(1):69–81, 1998.

[11] James J. Ryan. Teacher judgments of test items properties. *Journal of Educational Measurement*, 5(4):301–306, 1968.