XV JENUI. Barcelona, 8-10 de julio de 2009
ISBN: 978-84-692-2758-9
http://jenui2009.fib.upc.edu/

Estudio comparativo de diversos métodos de evaluación

José Luis Pérez de la Cruz
Dpto. LCC
Universidad de Málaga
Bulevar Luis Pasteur s/n
29017 Málaga
perez@lcc.uma.es

Eva Millán
Dpto. LCC
Universidad de Málaga
Bulevar Luis Pasteur s/n
29017 Málaga
perez@lcc.uma.es

Resumen

Es frecuente que en cada asignatura se empleen diversos métodos de evaluación (tests, examen de problemas, trabajos de curso). En este estudio, para varias asignaturas de Ingeniería Informática, se analizan estadísticamente los resultados obtenidos por los estudiantes mediante cada uno de los diversos métodos empleados.

1. Introducción

En la práctica cotidiana de la docencia universitaria de la informática se emplean diversos métodos para la evaluación del alumnado [3], [2]. Elegir uno u otro no es cuestión sin importancia: por ejemplo, podemos suponer que tendrá consecuencias sobre la motivación del alumnado y sobre la carga del trabajo del profesorado. Además, desde el punto de vista de la didáctica no hay un consenso claro sobre cuál sea "el mejor" método de evaluación; como se señala en [1], "existen (incluso entre los especialistas de la evaluación educativa) partidarios y detractores de cualquier formato de preguntas."

Los métodos de evaluación o pruebas se pueden clasificar en *pruebas de respuesta construída* y *de respuesta seleccionada*. Entre las primeras [6] tenemos los ensayos o preguntas largas, las preguntas cortas, las entrevistas y la evaluación de portafolios [4]. Entre las segundas tenemos las diversas variantes de los llamados habitualmente "tests". En el ámbito de las disciplinas científico-tecnológicas, dentro de las pruebas de preguntas largas tienen especial importancia los "exámenes de problemas", que en el caso de la informática, cuando consisten en la generación de código, suelen realizarse en el ordenador. En el ámbito de la informática, dentro de los portafolios tienen especial importancia los proyectos de desarrollo e implemantación de una aplicación.

En las diversas materias de la titulación "Ingeniero en Informática" de nuestra universidad se emplean todos los métodos mencionados. En este trabajo trataremos de aportar algunos datos y análisis empíricos para responder a la siguiente pregunta: cuando para evaluar al alumno en una asignatura empleamos un procedimiento compuesto por la aplicación de varios métodos diferentes, ¿arrojan todos ellos resultados análogos? Las respuestas obtenidas podrían servir de guía a los profesores a fin de depurar y simplificar los procedimientos de evaluación.

Para ello, hemos seleccionado tres asignaturas troncales u obligatorias de 2^o ciclo: "Inteligencia Artificial", "Procesadores de Lenguajes" e "Ingeniería del Conocimiento". Las tres son anuales, lo cual hace que los métodos de evaluación que en ellas se emplean sean complejos y consten de diversas pruebas.

Todos los análisis estadísticos que presentamos en este trabajo han sido realizados mediante el programa *R* [5], que se distribuye como software libre.

2. Caso 1: Inteligencia Artificial. Test y examen en laboratorio

Esta asignatura se imparte en 4^o curso y tiene asignados 10,5 créditos. Los objetivos fijados

```
16.- Tras
(DEFCLASS A ()  ((S1 :INITFORM 1 :READER S1)))
(DEFCLASS B (A) ((S1 :INITFORM 2 :READER S1)))
tenemos que (S1 (MAKE-INSTANCE 'B)) =>
a) 1
b) 2
c) Error
```

Figura 1: Ejemplo de pregunta de test de Inteligencia Artificial

para el estudiante son:

1. Asimilación de los conceptos y técnicas básicas de la IA desde el punto de vista teórico.

2. Aplicación práctica de lo anterior mediante su implementación en un lenguaje de alto nivel.

3. Aproximación a los sistemas de percepción.

Se sigue un método de evaluación compuesto. Concretamente, para evaluar la consecución del objetivo (2), se realizan dos pruebas:

- Una prueba de respuesta seleccionada ("test") referente a la sintaxis y semántica del lenguaje de alto nivel empleado (Common Lisp;. un ejemplo de pregunta aparece en el cuadro 1). Llamemos T a la calificación obtenida (0-10).

- El diseño e implementación, en tiempo y condiciones controladas, de la solución programada a un problema planteado ("laboratorio") Llamemos L a la calificación obtenida (0-3,5).

Se han analizado los datos correspondientes al curso 2007/08. Hubo un total de 40 alumnos que fueron evaluados en la convocatoria de junio tanto en laboratorio como en el test. La distribución de las calificaciones se muestra en la figura 2 y se resumen en la tabla 1. En esta misma tabla se muestran sus varianzas y covarianzas. Por último, si consideramos únicamente 2 valores para cada variable (menor que la media, o mayor-igual que la media), se obtienen los valores de la parte inferior de la tabla.

	L	T
Min	0.010	1.500
Mediana	2.750	7.000
Media	2.284	6.594
Max	3.500	9.500

varianza	L	T
L	1.198	0.860
T	0.860	3.717

	$T \leq \overline{T}$	$T > \overline{T}$
$L \leq \overline{L}$	10	9
$L > \overline{L}$	7	14

Cuadro 1: Resumen de las calificaciones de IA

Sustituyendo L por $10L/3{,}5$, a fin de que su rango sea también 0–10, el ajuste lineal sin término independiente viene dado por

$L = 0{,}9645T$

El ajuste lineal con término independiente viene dado por

$L = 2{,}1690 + 0{,}6608T$

Para estimar si ambas variables están correlacionadas o no, calcularemos el valor del coeficiente ρ de Pearson. Como es sabido, si $\rho = 0$ no existe ninguna correlación lineal, pero si $\rho = 1$ el ajuste lineal es perfecto. En este caso, el coeficiente de correlación de Pearson es $\rho = 0{,}407$, lo cual indica cierta correlación positiva. Además, su intervalo de confianza al 95 % es [0.110, 0.638], por lo que podemos asegurar con un 95 % de probabilidad que existe una correlación positiva entre ambas variables, al menos de 0,110.

Podemos esquematizar como sigue las conclusiones extraídas de este análisis:

- Las calificaciones del test y del examen de laboratorio *sí* están correlacionadas positivamente, por lo que podemos afirmar que hasta cierto punto miden la misma magnitud, o magnitudes relacionadas.

- El mejor ajuste lineal que se obtiene sin término independiente nos dice que, prác-

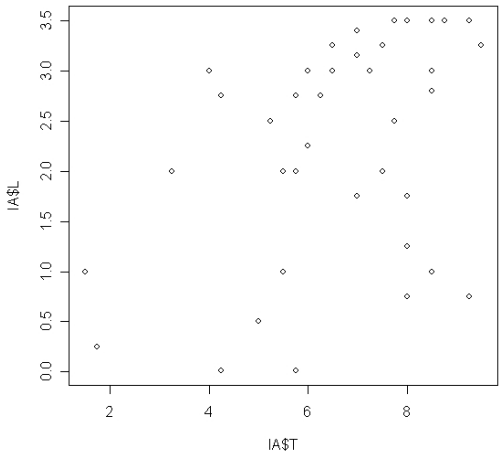

Figura 2: Calificaciones de IA

ticamente, $L = T$. Sin embargo, si admitimos término independiente, éste no resulta ser 0, sino aprox. 2. Es decir, parece que independientemente de la nota obtenida en el test, los alumnos obtienen un mínimo de 2 ptos sobre 10 en el examen de laboratorio.

- Si se hubieran establecido notas mínimas o de corte para el test, algunos alumnos que superaron el examen de laboratorio no habrían sido tenidos en cuenta. Por ejemplo, 2 alumnos obtuvieron menos de 5 puntos sobre 10 en el test y sin embargo obtuvieron más de 2.5 puntos sobre 3.5 en el laboratorio.

Por tanto, desde el punto de vista práctico:

- Aunque ambas pruebas miden magnitudes muy relacionadas, no es posible prescindir de una de ellas sin perder información
- Si se pretende que ambas pruebas evalúen conjuntamente las habilidades de programación, sería injusto establecer notas "de corte" en el test, ya que hay casos en los que en el laboratorio se demuestra notable pericia pese a haber fallado en la prueba de respuesta seleccionada. Nótese que las posibles perturbaciones en las prestaciones del alumno en el test se ven muy atenuadas por el hecho de que el alumno tiene tres oportunidades no excluyentes de presentarse al mismo; si no lo supera en febrero, puede hacerlo en mayo, y si aún no lo supera, puede hacerlo en junio.

3. Caso 2: Procesadores de Lenguajes. Test y examen en laboratorio

Esta asignatura se imparte en 4º curso y tiene asignados 10,5 créditos. Los objetivos fijados para el estudiante son:

1. Estudio teórico de los fundamentos de la compilación y las técnicas más utilizadas.
2. Análisis, diseño e implementacion de procesadores de lenguajes de programación mediante las herramientas estudiadas.

Se sigue un método de evaluación compuesto. El camino recomendado y más habitualmente seguido por los estudiantes ("evaluación por curso") es el siguiente:

- A lo largo del curso se realizan cuatro pruebas de respuesta seleccionada T_1, T_2, T_3, T_4, con calificaciones de 0 a 1,5, tras el estudio cada bloque temático (en la figura 3 se muestra uno de los ítems presentados en estas pruebas). Llamemos $T = 10\Sigma T_i/6$, de forma que el rango de T será de 0 a 10.
- En el examen final de junio se realiza la prueba P, consistente en el diseño e implementación, en tiempo y condiciones controladas, de (un fragmento de) un pequeño compilador. Para P también emplearemos en este estudio un rango de 0 a 10.

Se han analizado los datos correspondientes al curso 2007/08. Hubo un total de 56 alumnos que se presentaron al examen de laboratorio en la convocatoria de junio. La distribución de las calificaciones se muestra en la figura 4 y se resume en la tabla 2. Nótese que todas las

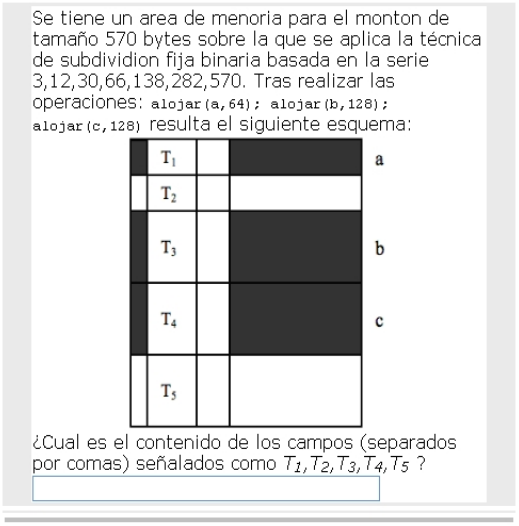

Figura 3: Ejemplo de pregunta de test de Procesadores de Lenguajes

	T	P
Min	1.875	0.000
Mediana	6.746	5.000
Media	6.398	3.449
Max	9.062	10.000

varianza	T	P
T	2.299	2.344
P	2.344	12.045

	$T \leq \overline{T}$	$T > \overline{T}$
$P \leq \overline{P}$	18	8
$P > \overline{P}$	10	20

varianza	T	P
T	1.515	0.733
P	0.733	2.940

Cuadro 2: Resumen de las calificaciones de PL

$P < 5$ se han almacenado como $P = 0$, lo cual introduce una cierta perturbación en los datos. Las varianzas y covarianzas también se muestran en esta tabla. De nuevo señalamos que los datos (relativamente) espurios con $P = 0$ hacen que estas varianzas no sean las reales de la población.

Si consideramos únicamente 2 valores para cada variable (menor que la mediana, o mayor-igual que la mediana), obtenemos los valores igualmente mostrados en la tabla 2.

El ajuste lineal sin término independiente viene dado por

$P = 0{,}564T$

El ajuste lineal con término independiente viene dado por

$P = -3{,}075 + 1{,}100T$

El coeficiente de correlación de Pearson es $\rho = 0{,}445$. Su intervalo de confianza al 95 % es [0.207, 0.634].

Si repetimos el análisis únicamente con los datos de los alumnos con $5 \leq P$ (figura 5, obtenemos por último las varianzas de la parte inferior de la tabla 2.

El ajuste lineal con término independiente viene dado por

$P = 3{,}074 + 0{,}484T$

El coeficiente de correlación de Pearson es $\rho = 0{,}374$. Su intervalo de confianza al 95 % es [-0.015, 0.629]. Por tanto, y debido a lo reducido de la muestra, no podemos rechazar a este nivel que $\rho = 0$

Podemos esquematizar como sigue las conclusiones extraídas de este análisis:

- Las calificaciones del test y del examen de laboratorio están correlacionadas positivamente. Este resultado resulta un tanto inesperado, dado que pretenden medir competencias diferentes. Podría pensarse que la teoría es prerrequisito del laboratorio, pero ello no parece ser así: nótese que el número de alumnos con T bajo y P alto es comparable al número de alumnos con T alto y P bajo.
- Los ajustes lineales no resultan demasiado indicativos, debido al filtrado y prepocesamiento al que fueron sometidos los datos previamente a su almacenamiento.
- Pese a la correlación existente, de nuevo existe una minoría significativa de alumnos que se comportan de manera muy diferente en ambas pruebas.

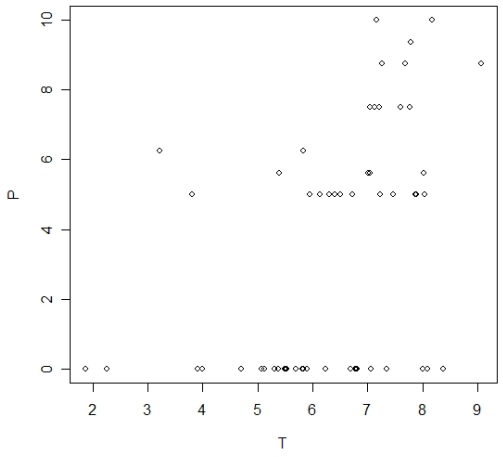

Figura 4: Calificaciones de PL

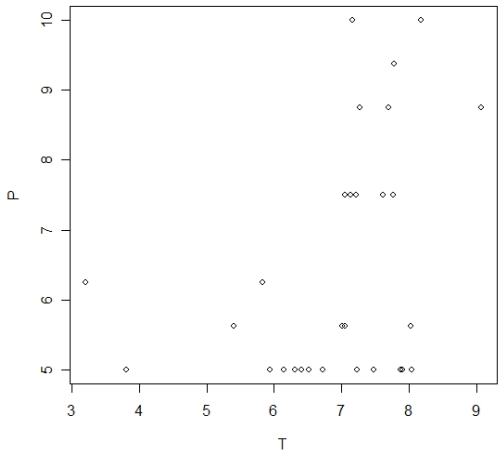

Figura 5: Calificaciones de PL ($5 \leq P$)

Por tanto, desde el punto de vista práctico:

- Aunque ambas pruebas están en principio destinadas a medir competencias diferentes, la calificación en una puede servir para predecir la calificación en la otra. Sin embargo, esta predicción no es demasiado exacta, así que resultaría necesario mantener ambas pruebas.

- Con carácter general, parece conveniente almacenar los datos de la manera más fiel posible, a fin de poderlos analizar posteriormente.

4. Caso 3: Ampliación de Ingeniería del Conocimiento. Examen de problemas y proyecto

Esta asignatura se imparte en 5º curso y tiene asignados 9 créditos. Los objetivos fijados para el estudiante son:

1. Poseer y comprender conocimientos sobre las técnicas formales de representación y manejo del conocimiento.

2. Manejar eficientemente herramientas para la implementación de sistemas basados en el conocimiento (SBC).

3. Desarrollar de un pequeño SBC.

Para evaluar la consecución del objetivo (1), se realizan a lo largo del curso dos pruebas de respuesta construída (problemas), referentes a las técnicas expuestas en cada cuatrimestre. Llamemos A_1 y A_2 a sus calificaciones numéricas (0–10) y $A = (A_1 + A_2)/2$

Para evaluar la consecución de los objetivos (2) y (3), se evalúa un portafolio formado por un proyecto o trabajo que cada alumno (individualmente) debe realizar a lo largo del segundo cuatrimestre. Este trabajo consiste en la implementación de un pequeño SBC sobre un tema seleccionado por el mismo alumno y aprobado por los profesores. Por ejemplo, un proyecto típico puede ser "Sistema para seleccionar ayudas al desempleo".

Los rasgos evaluados son: la originalidad del dominio de aplicación y complejidad y dificultad del proceso de extracción del conocimiento; la amplitud del dominio de aplicación; las técnicas de programación empleadas; las técnicas derazonamiento empleadas; el diseño de la interfaz de usuario y/o integración con otros programas; y la calidad de la documentación. Llamemos B a la calificación numérica resultante (0–10).

Se han analizado los datos correspondientes al curso 2007/08. Hubo un total de 45 alumnos que se presentaron a los exámenes y entregaron los trabajos en la convocatoria de junio. La distribución de las calificaciones se muestra resumida en la tabla 3. Sus varianzas y covarianzas se muestran en la misma tabla.

La distribución de las calificaciones se muestra en las figuras 6 (calificaciones A_1 y A_2) y 7 (calificaciones A y B).

Si consideramos únicamente 2 valores para A y B (menor que la media, o mayor-igual que la media), obtenemos los valores de la tabla 3 (parte central).

Por último, si consideramos únicamente 2 valores para A_1 y A_2 (menor que la media, o mayor-igual que la media), obtenemos los valores de la tabla 3 (parte inferior).

Los ajustes lineales sin términos independientes vienen dados por

$A_1 = 0{,}839A_2$

$B = 0{,}9645A$

El coeficiente de correlación de Pearson entre A_1 y A_2 es $\rho = 0{,}603$. Su intervalo de confianza al 95 % es [0.377, 0.762]. La correlación es bastante significativa.

El coeficiente de correlación de Pearson entre A y B es $\rho = -0{,}017$. Su intervalo de confianza al 95 % es [-0.309, 0.278]. No podemos pues afirmar que exista ninguna correlación.

Podemos esquematizar como sigue las conclusiones extraídas de este análisis:

- Las calificaciones del proyecto y de los exámemes de problemas no presentan ninguna correlación significativa. Por tanto, miden magnitudes diferentes.
- Sin embargo, las calificaciones de ambos exámenes si presentan cierta correlación.

	A_1	A_2	A	B
Min	1.670	1.170	1.420	5.00
Mediana	6.670	7.830	7.000	7.25
Media	6.381	7.544	6.963	7.30
Max	10.000	9.250	9.250	9.75

varianza	A_1	A_2	A	B
A_1	2.716	1.497	2.105	0.011
A_2		2.266	1.882	-0.076
A			1.993	-0.032
B				1.826

	$A \leq \overline{A}$	$A > \overline{A}$
$B \leq \overline{B}$	12	10
$B > \overline{B}$	10	13

	$A_1 \leq \overline{A_1}$	$A_1 > \overline{A_1}$
$A_2 \leq \overline{A_2}$	15	7
$A_2 > \overline{A_2}$	7	16

Cuadro 3: Resumen de las calificaciones de AIC

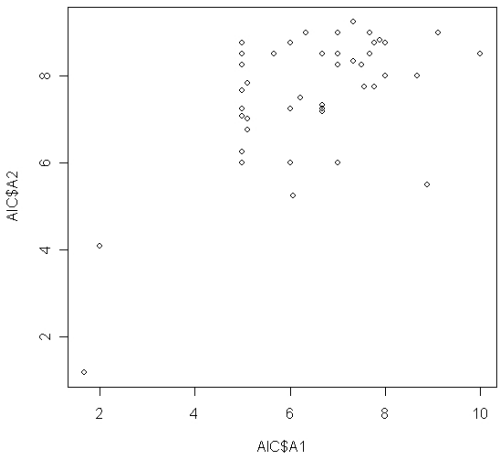

Figura 6: Calificaciones de AIC (exámenes)

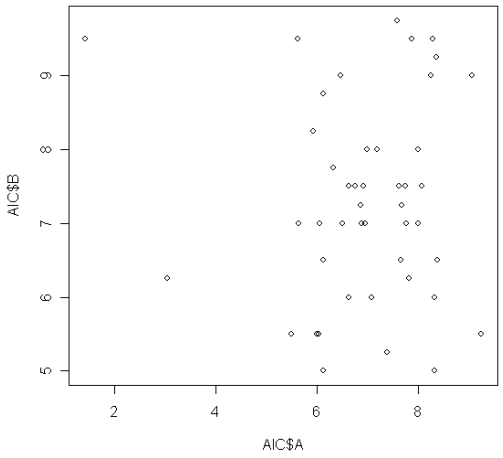

Figura 7: Calificaciones de AIC (exámenes y proyecto)

> Podríamos pensar que dependen en cierta medida de una única habilidad genérica del alumno.

Por tanto, desde el punto de vista práctico, no parece conveniente prescindir de ninguno de los dos criterios analizados.

5. Conclusiones

En este trabajo se ha analizado para tres asignaturas de 2^o ciclo de Ingeniero en Informática la relación existente entre los diversos métodos de evaluación empleados en cada una de ellas.

Como resultado de este análisis, podemos afirmar que los resultados arrojados por los diversos métodos, aun teniendo a veces cierta correlación, en ningún caso llegan a coincidir tanto que sea posible prescindir de uno de ellos sin pérdida significativa de información, lo que quizás coincida con la intuición que manifiestan la mayoría de los profesores experimentados.

Sin embargo, en la práctica cotidiana vemos también a menudo que algunos compañeros realizan una evaluación basada sólo un un método (típicamente, tests).

A la luz de nuestros resultados, parecería prudente recomendar a todos los docentes que empleen siempre un repertorio de métodos de evaluación suficientemente variados.

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 2009-02-10.

[2] Daniel González Morales, José Luis Roda García, and Luz Marina Moreno de Antonio. Aplicando diferentes técnicas de evaluación. In *Actas de las XIV Jornadas de Enseñanza Universitaria de Informática, Jenui 2008*, pages 403 – 410, Granada, Julio 2008.

[3] Isabel Guitart, M. Elena Rodríguez, Jordi Cabot, and Montse Serra. Elección del modelo de evaluación: caso práctico para asignaturas de ingeniería del software. In *Actas de las XII Jornadas de Enseñanza Universitaria de Informática, Jenui 2006*, pages 191 – 198, Bilbao (Vizcaya), Julio 2006.

[4] María A. Pinar Sepúlveda and Joaquín Gracia Morán. Aplicación del portafolio como estrategia de evaluación formativa. In *Actas de las XIII Jornadas de Enseñanza Universitaria de Informática, Jenui 2007*, pages 249 – 256, Teruel, Julio 2007.

[5] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[6] J. Alonso Tapia. Evaluación del conocimiento: propósito, criterios, contexto y problemas. In J. Alonso Tapia, editor, *Evaluación del conocimiento y su adquisición. Volumen I: Ciencias Sociales*, pages 19–60. Ministerio de Educación y Cultura, 1997.