Análisis estadístico del uso de cuestionarios on-line como herramienta de evaluación.

A. Bia Platas; X. Barber i Vallés

Centro de Investigación Operativa
Universidad Miguel Hernández de Elche

RESUMEN

¿A quién no le gustaría poder montar cuestionarios fácilmente, aplicarlos rápidamente y obtener automáticamente una puntuación que midiese de forma objetiva el conocimiento y las habilidades adquiridas por los alumnos? Pero la experiencia nos dice que hacer buenos cuestionarios no es tarea fácil, y cuando lo es, los cuestionarios no son buenos. Muchas veces nos preguntamos: ¿sirven los cuestionarios online como método de evaluación? ¿Qué condiciones se deben cumplir para aplicarlos efectivamente? ¿y como herramienta de aprendizaje con realimentación? Un test puede resultar fácil para un grupo de personas, y resultar difícil para otro. No hay forma de saberlo de antemano, pero a posteriori se pueden analizar los resultados estadísticamente y sacar conclusiones sobre la eficacia de este método. En este artículo analizamos los posibles usos de los cuestionarios online, así como su efectividad como método de evaluación en diferentes escenarios, y presentaremos algunos casos reales analizados mediante técnicas estadísticas simples. Discutiremos también los problemas encontrados en su aplicación a distancia y presencial, y presentaremos algunas recomendaciones para su construcción y aplicación eficaz.

Palabras clave: cuestionarios online, evaluación, eLearning, estadística, docencia no presencial, Moodle.

1. INTRODUCCIÓN

El principal objetivo de un cuestionario tipo test es **evaluar el aprendizaje** ("summative assessment" [Gov2002]), obteniendo una calificación numérica que sirva como medida de los conocimientos adquiridos. Construir cuestionarios tipo test que sirvan para este cometido, no es tarea fácil. Por ejemplo, un test muy fácil y con un tiempo excesivo, no serviría para este propósito.

El objetivo secundario es como **herramienta didáctica**, para **reforzar el aprendizaje** ("formative assessment" [Gov2002]). Éste es el caso de los test con retroalimentación inmediata, con ayudas y sugerencias tras los fallos, y a veces con la posibilidad de múltiples reintentos.

Estos dos objetivos, evaluación y reforzamiento del aprendizaje, no son compatibles entre sí. Un buen test de evaluación no puede ofrecer una realimentación inmediata, ayudas y sugerencias, ni la posibilidad de reintentos múltiples. Esto es especialmente cierto cuando el test online se mantiene abierto por un período de tiempo prolongado, como suele hacerse en algunos cursos a distancia, para mayor comodidad de los estudiantes que tienen distintos horarios y obligaciones. Una solución de compromiso consiste en dar la realimentación una vez cerrado el test, cuando ya nadie más puede hacerlo. Govindasamy no recomienda los test como único método de evaluación, pero destaca su gran potencial como herramienta formativa: "un MCQ[i] diseñado cuidadosamente puede ayudar a los estudiantes a adquirir un conocimiento profundo del contenido" [Gov2001], y sugiere el uso del cuestionario en combinación con las búsquedas electrónicas de las respuestas como técnica formativa y de profundización.

El tercer objetivo de los test es **hacer que los alumnos estudien**, ya que sin la presión de tener que someterse a pruebas de evaluación y de cumplir plazos, muchos de ellos no estudiarían.

En el presente trabajo, basándonos en una colección de datos de test realizados en un curso online con más de un centenar de alumnos de toda España, hemos intentado, mediante la aplicación de métodos estadísticos, responder a preguntas como las siguientes:

- ¿Cuándo un test resulta demasiado fácil, o demasiado difícil?
- ¿Conviene dar retroalimentación inmediata o diferida?
- ¿Cuál es el tiempo óptimo para realizar un test?

- ¿Debemos limitar el tiempo o dar un tiempo holgado?
- ¿Hay alguna forma de saber si ha habido *mala praxis* por parte de los alumnos? (p. ej., pasarse las respuestas, o responder el test consultando el material de estudio)

Literatura sobre el tema

Hay muchos artículos que tratan de la evaluación de la educación a distancia (eLearning) [Lev2003, Att2006], que discuten sus bondades y desventajas de forma general, pero que dicen poco sobre la evaluación del propio aprendizaje del alumno en un entorno online. Otros son más específicos sobre la evaluación usando cuestionarios tipo test [Sch2009, Thu2010]. Entre la gran cantidad de información general y guías de buenas prácticas sobre eLearning que hay en la Web, queremos destacar un curso online llamado "e-Learning for Development" del Rice Knowledge Bank[ii], que trata todos los aspectos de la creación de un proyecto de docencia online no presencial.

Un artículo muy interesante de T. Govindasamy [Gov2002], le da mucha importancia a la evaluación, como reforzador de los métodos de aprendizaje que adopta el alumno. Dice también que si al alumno se lo evalúa en habilidades de alto nivel, es probable que adopte un enfoque holístico profundo respecto al e-Learning, mientras que si se le evalúa en habilidades de bajo nivel, es probable que practique el "indeseable enfoque atómico superficial" mencionado por Twomey [Two1996]. Habla también de malas prácticas en la evaluación, agravadas en los entornos de e-Learning, donde los docentes que se ven abrumados por su nuevo rol en la docencia virtual son "propensos a sucumbir a la conveniencia del uso de cuestionarios calificados automáticamente" como método de evaluación.

2. METODOLOGÍA

El problema que nos hemos planteado es estudiar un conjunto de 8 cuestionarios tipo test (identificados a los efectos del estudio por número de asignatura-bloque, como A1B1… A3B3) que se aplicaron a 170 alumnos de diferentes partes de España en un curso online[iii] a lo largo de un curso académico. En todas las ocasiones, hubo algunos alumnos que no hicieron el test. Por ello las cifras de intentos son siempre menores que 170. De todos modos, los cuestionarios han sido realizados por un número bastante grande de alumnos, lo cual contribuye a la confiabilidad de los resultados del estudio.

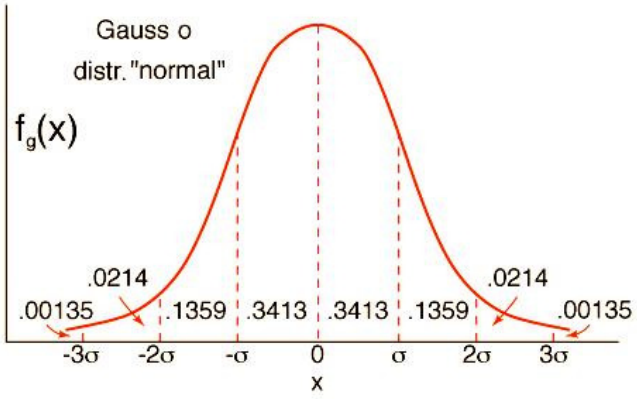
Los tests online, fueron aplicados a través de una plataforma Moodle, que ofrece muchas posibilidades de configuración relativas a la aplicación del test y de la

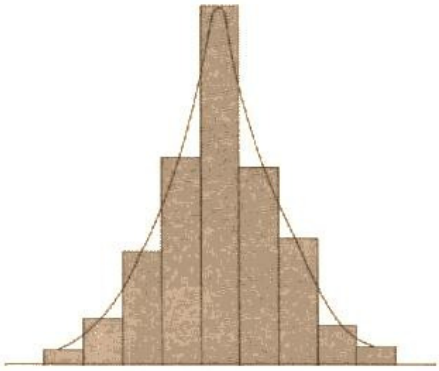
retroalimentación recibida por el alumno. La misma, permite obtener curvas de frecuencias de las notas de las pruebas (diagramas en rojo), y descargar los datos en diversos formatos para su post-procesamiento (p.ej., Excel, Open Document Format y texto con separadores). Para un análisis más profundo y meticuloso de es estos datos, hemos utilizado el entorno de computación estadística "R"[iv].

3. RESULTADOS

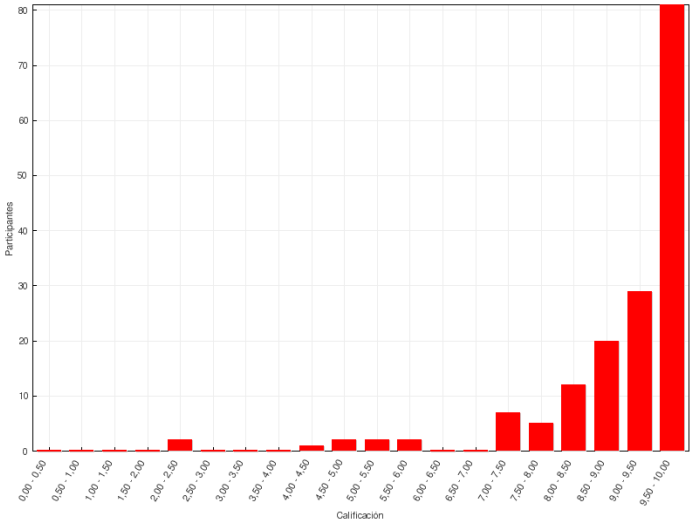
Test fácil o test difícil

Un test puede ser fácil para un grupo de personas, y resultar difícil para otro. No hay forma de saber si un test va a ser fácil o no, hasta que todos lo hayan hecho. Quizás la forma más rápida de ver si un test ha resultado fácil o difícil es mediante un gráfico de la distribución de las calificaciones obtenidas. Si el número de alumnos es lo suficientemente grande, la distribución de calificaciones suele parecerse, aproximadamente, a una campana de Gauss, la cual es una distribución ideal (ver figura, abajo a la izquierda). En realidad lo que obtenemos es una curva de frecuencias que se asemeja a una campana de Gauss (ver figura, abajo a la derecha). Las últimas versiones de la plataforma de docencia online Moodle, provee automáticamente estas gráficas (gráficas en rojo, más abajo).



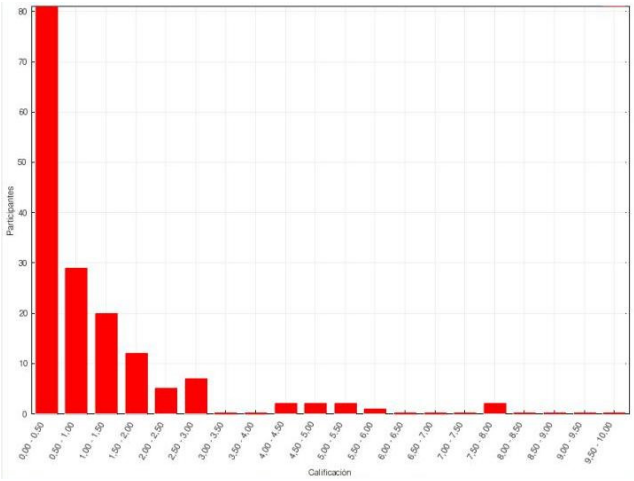

Curva de frecuencias

Por ejemplo, el test A2B1 fue muy fácil. Lo hicieron 164 alumnos y la calificación promedio fue de 9,02 puntos. Tenía sólo 15 preguntas y se disponía de una hora para hacerlo, lo cual daba tiempo de buscar las respuestas en los apuntes. En este caso, la curva esta tan desplazada a la derecha que la mitad derecha de la campana de Gauss no existe.



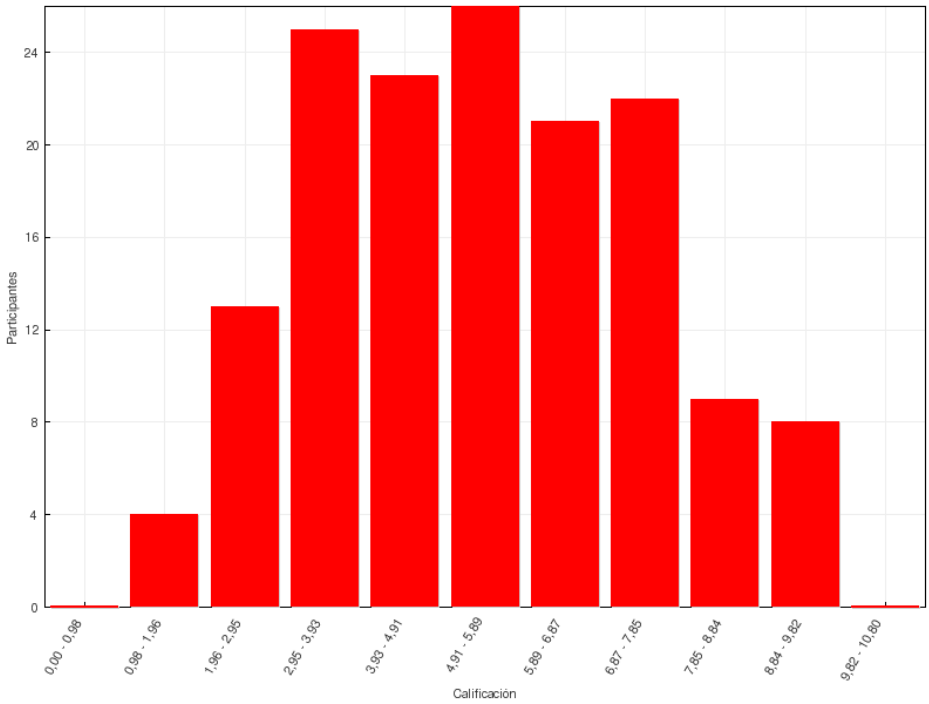
Test muy fácil

Por el contrario, un test muy difícil, sería lo opuesto al anterior:

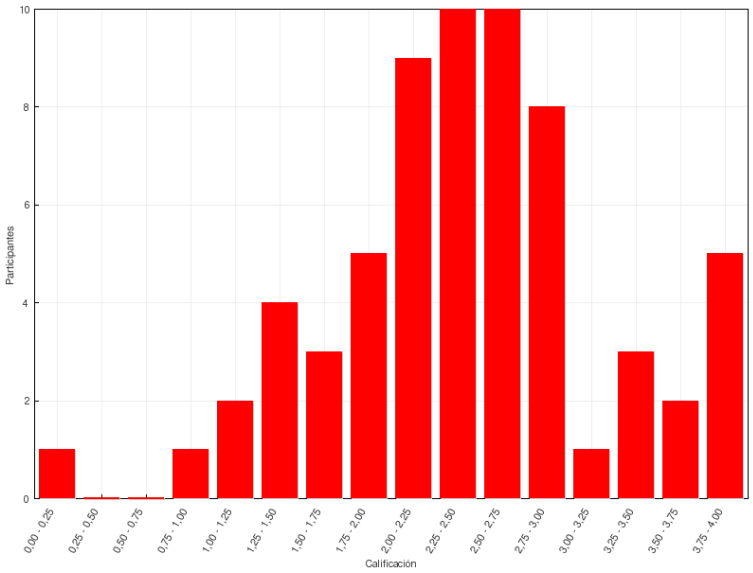


Test muy difícil

Otro de los cuestionarios, el test A3B2, fue de dificultad media, o media-alta. Lo hicieron 151 alumnos, y la calificación media fue de 5,35 puntos. En este caso vemos que el gráfico se parece ya más a una campana de Gauss completa:



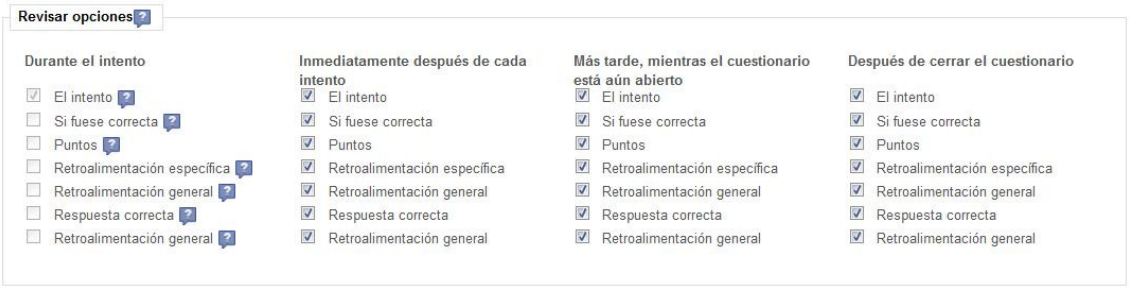
El test de recuperación A3B2r ("repesca"), lo hicieron sólo 64 personas (los que habían sacado menos de 5 puntos en el test A3B2). La calificación media fue 6,05 puntos. Se obtuvieron mejores resultados, y la curva resultante se ve bastante más a la derecha que la de A3B2. Se trata de una mejora significativa si tenemos en cuenta que estos alumnos estaban en la mitad izquierda de la curva anterior.



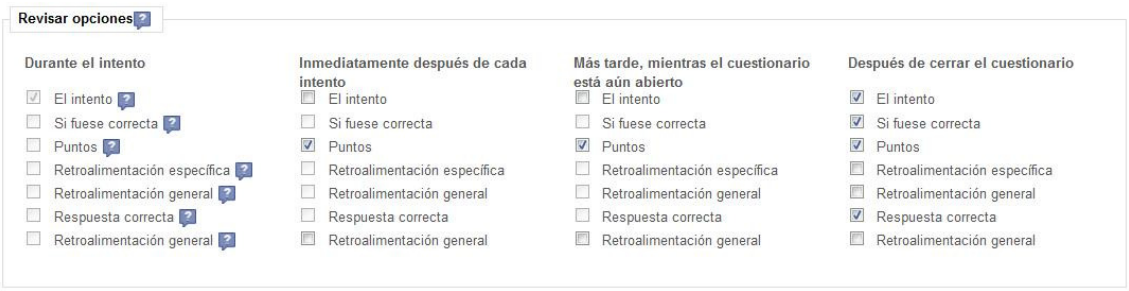
Cuestionarios restrictivos y permisivos

A lo largo del curso, se realizaron dos tipos de test, que hemos clasificado como **permisivos** y **restrictivos** respectivamente. En los **test permisivos** se les dio más tiempo por pregunta que en los restrictivos, y se les permitía ver las respuestas correctas inmediatamente tras terminar el test. El tiempo asignado les daba tiempo suficiente para

buscar las respuestas de las preguntas de las que no estuviesen seguros en el material de estudio. Además, al tratarse de un curso online y a distancia, el cuestionario permanecía abierto durante varios días, e incluso semanas. Existía entonces la posibilidad de que los alumnos que lo hicieran primero tomasen nota de las respuestas correctas y se las pasasen a otros alumnos antes de que éstos hicieran su cuestionario.



En los **test restrictivos**, se modificaron los ajustes del Moodle para dificultar este tipo de mala praxis, dándoles menos tiempo por pregunta, y postergando la retroalimentación hasta que el cuestionario quedase cerrado completamente y nadie más lo pudiese contestar. De este modo se hace más difícil el poder copiar las preguntas, y más aún las respuestas correctas.



Calificaciones

	TIPO	RESPUESTAS	NOTA MEDIA
A1B1	Permisivo	169	7.9±1.5
A1B2	Permisivo	169	8.9±1.1
A1B3	Permisivo	166	9.4±1
A2B1	Permisivo	165	9±1.4
A2B2	Permisivo	163	8.7±1.5
A3B2	**Restrictivo**	151	5.4±2
A3B2 REPESCA	**Restrictivo**	64	6.1±1.9
A3B3	**Restrictivo**	150	9.4±0.7

Lo primero que salta a la vista es en **gran descenso en la calificación media en el primer test en que no se les proporcionaba las respuestas correctas al finalizar el**

mismo (retroalimentación diferida). Si estudiamos los percentiles de los test, encontramos resultados todavía más curiosos.

	TIPO	MÍN	P5	P10	P20	P25	MEDIANA	P75	P90	MÁX.
A1B1	Permisivo	2	6	6	7	7	8	9	10	10
A1B2	Permisivo	3	7	7	8	8	9	9	10	10
A1B3	Permisivo	5	8	8	8	9	10	10	10	10
A2B1	Permisivo	2	6	8	8	9	9	10	10	10
A2B2	Permisivo	3	5	7	8	8	9	10	10	10
A3B2	**Restrictivo**	1	2	3	4	4	5	7	8	10
A3B2 REPESCA	**Restrictivo**	1	3	1	4	5	6	7	9	10
A3B3	**Restrictivo**	5	8	9	9	9	10	10	10	10

Analizando los permisivos, se observa como en el grupo de test A1, el 95% de los alumnos tienen una nota superior al 6 y que el 50% tanto del A1 como del A2 superan el 8. Esto hace pensar que estos test (grupos A1 y A2) eran excesivamente fáciles a los efectos de mostrar las diferencias de aprendizaje entre los alumnos, y ello puede venir dado por diferentes causas, o bien un tiempo excesivo o bien que ha habido una mala praxis entre los alumnos (p.ej. pasarse información de las preguntas entre ellos).

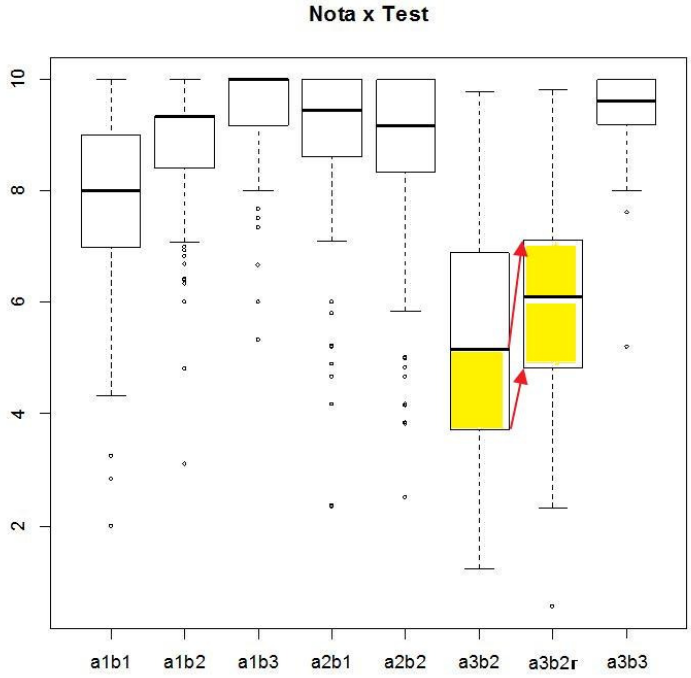
Los tiempos para cada uno de los test son los siguientes:

	TIPO	TIEMPO	NRO.PREG.	T/PREG
A1B1	Permisivo	60’	10	6’
A1B2	Permisivo	90’	15	6’
A1B3	Permisivo	30’	5	6’
A2B1	Permisivo	40’	15	2’40’’
A2B2	Permisivo	20’	12	1’40’’
A3B2	**Restrictivo**	45’	22	2’3’’
A3B2 REPESCA	**Restrictivo**	50’	22	2’16’’
A3B3	**Restrictivo**	40’	25	1’36’’

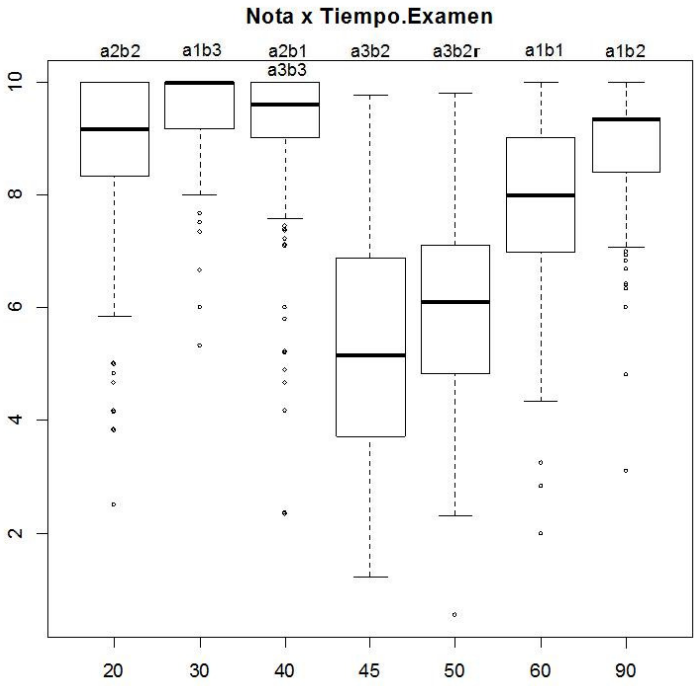
Si a esta tabla le unimos valores de los percentiles anteriores, se observa claramente cómo el tiempo por pregunta no discrimina tanto como lo que se pudiera esperar. En la tabla siguiente comprobamos cómo las variaciones no se deben al tiempo por pregunta, sino que vienen dadas por otros factores.

T/PREGUNTA	MEDIA±DESV.	MÍN	P25	MEDIANA	P75	MÁX.
1’ 36’’	9.4±0.7	5	9	10	10	10
1’ 40’’	8.7±1.5	3	8	9	10	10
2’ 03’’	5.4±2	1	4	5	7	10
2’ 16’’	6.1±1.9	1	5	6	7	10
2’ 40’’	9±1.4	2	9	9	10	10
6’ 00’’	8.7±1.4	2	8	9	10	10

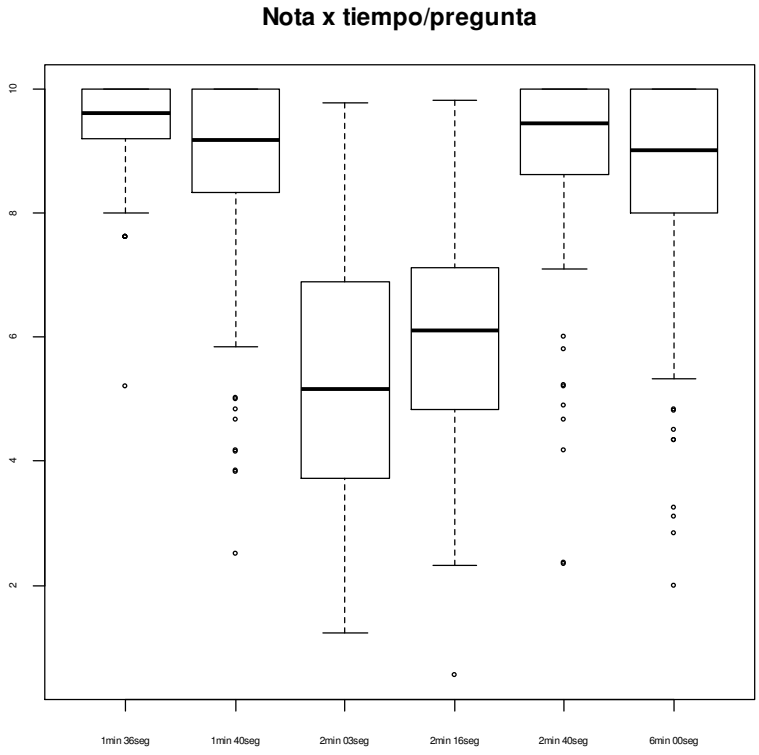
Veamos gráficamente estos resultados para ver las distribuciones por prueba, tiempo, y ratio tiempo (tiempo empleado sobre tiempo total de la prueba).



En estos gráficos de cajas se puede ver la distribución de la nota para los diferentes test. Se puede ver como existen ligeras diferencias en el comportamiento de las respuestas, y como claramente **el test A3B2 resultó ser el más difícil** y que aquellos que no lo superaron, cuando tuvieron una segunda oportunidad (test de repesca A3B2r), sí lo superaron y también subieron la nota media.

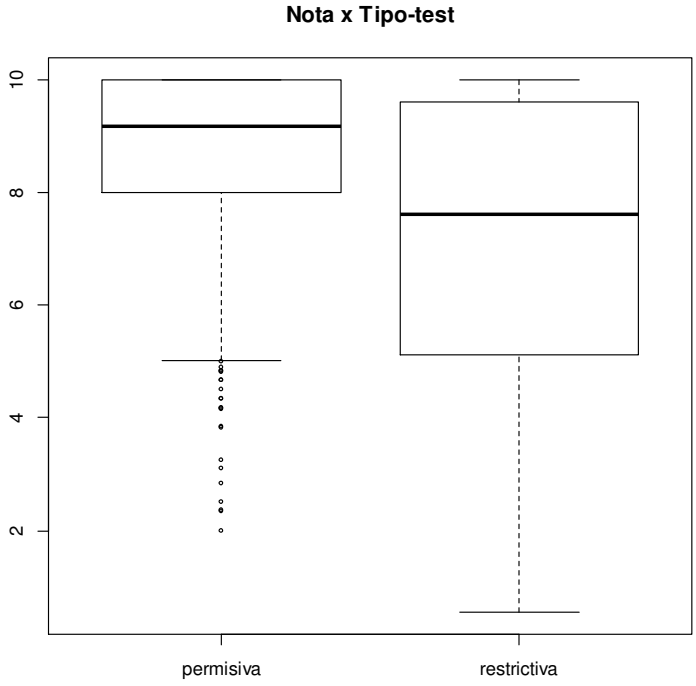


Cuando se analiza la nota obtenida según el tiempo total del test, independientemente del número de preguntas que este tenga, podemos ver si existe o no fatiga mental a la hora de responder. En este caso vemos que la duración de la prueba no afecta la calificación de forma significativa, pues podemos reconocer los mismos patrones de antes.

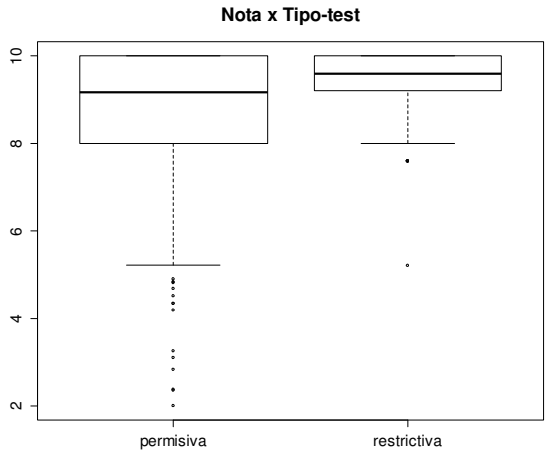


Respecto a tener en cuenta el **tiempo que tiene el alumno por pregunta**, el patrón se sigue conservando y no parece afectar el hecho de tener más o menos tiempo. Esto nos hace reflexionar que con **un tiempo entre 1'30'' y 2'00'' por respuesta es suficiente**.

Ahora bien, si analizamos lo mismo por tipo de test (permisivo o restrictivo), el grafico sí muestra diferencias, pero estas se deben al hecho de meter en el mismo saco el A2B2 y el A3B2r (repesca), y esto no sería del todo correcto.



Si sólo consideramos los test permisivos y el A3B3, el gráfico obtenido sería:

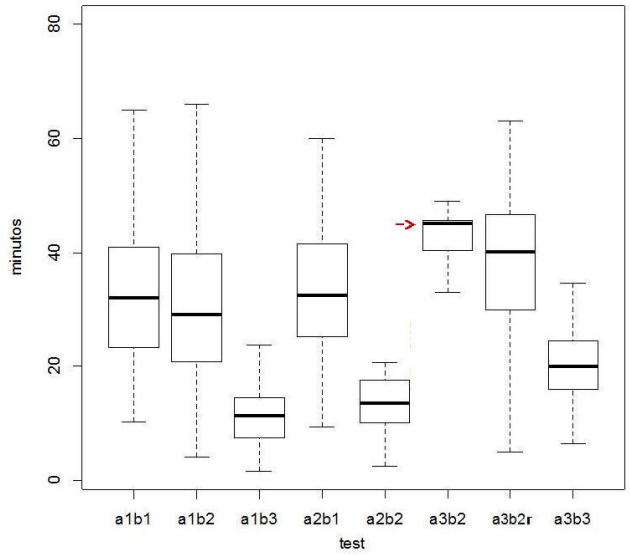


Y por tanto no hay diferencia por tipo de Test, a no ser que psicológicamente y vistos los resultados del test anterior se haya querido poner un tipo de test más fácil.

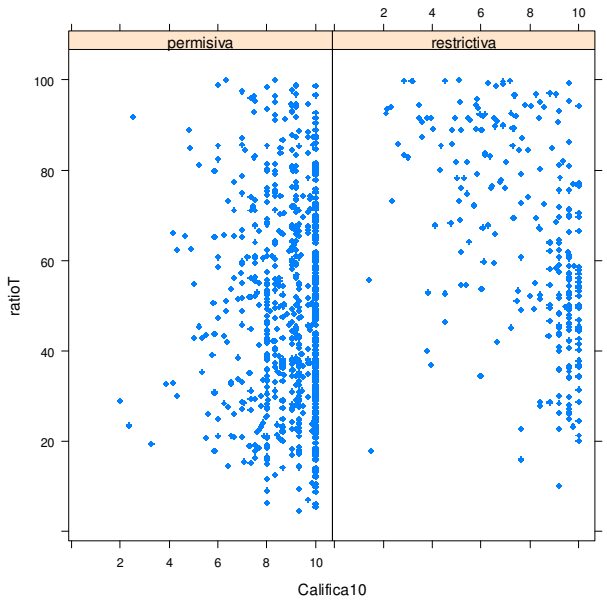
Tiempos para realizar las pruebas

Estudiaremos ahora los tiempos reales requeridos para completar los cuestionarios, y no el tiempo máximo permitido para hacerlos.

	TIPO	T. MÁX.	T.REAL MEDIO ±DESV.	NOTA MEDIA
A1B1	Permisivo	60’	32.7±12.6	7.9±1.5
A1B2	Permisivo	90’	31±13.9	8.9±1.1
A1B3	Permisivo	30’	11.7±5.5	9.4±1
A2B1	Permisivo	40’	33.1±11.6	9±1.4
A2B2	Permisivo	20’	13.5±4.8	8.7±1.5
A3B2	**Restrictivo**	45’	41.8±7.5	5.4±2
A3B2 REPESCA	**Restrictivo**	50’	37.9±11.5	6.1±1.9
A3B3	**Resticitivo**	40’	20.8±7.4	9.4±0.7

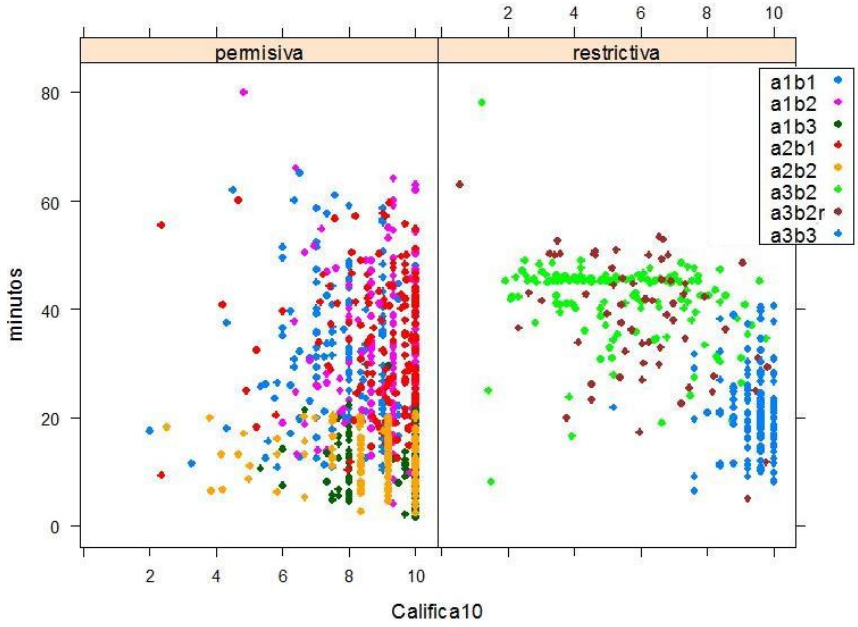


Nótese que en el test a3b2, que fue el test más difícil de los restrictivos, el tiempo se les quedó corto, mientras que en la repesca (a3b3r) que fue muy similar en dificultad, por estudiar más y tener la experiencia del test previo, no sólo mejoraron los resultados, sino que no estuvieron tan justos de tiempo.



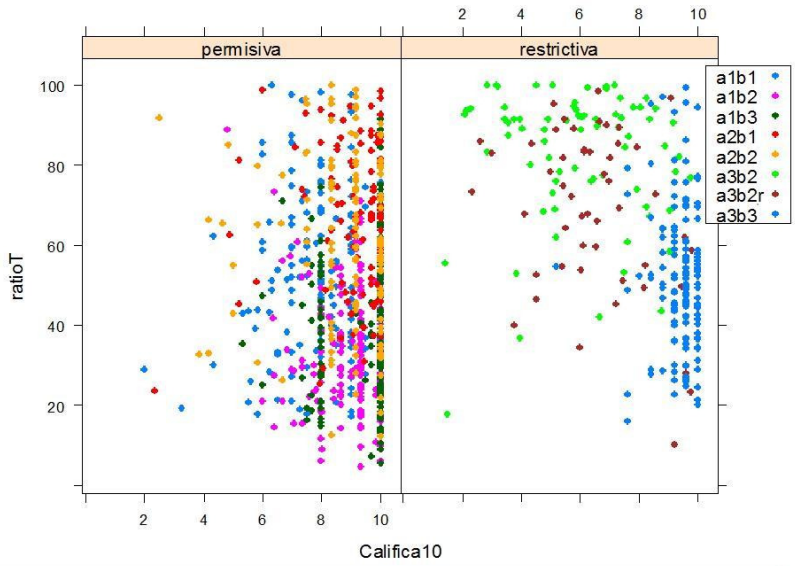
Numéricamente, existe una correlación negativa de -0.4 entre lo que se tarda en responder y la nota media, es decir **a mayor tiempo que el alumno tarda en responder el test, peor nota obtiene**. En las pruebas de tipo test es lo que vulgarmente se conoce como el "pescador de respuestas". Esta correlación es bastante moderada, y aunque significativa en el gráfico podemos ver como no hay un patrón definido entre notas y tiempos para realizar la prueba, a no ser que diferenciemos por otras variables, como hace el siguiente gráfico.

Si consideramos la variable **ratioT**, como el porcentaje del tiempo que tardan en realizar el test (minutos en realizar el Test/Tiempo maximo)*100, la correlación disminuye a -0.2, y muestra el gráfico de la parte de arriba.

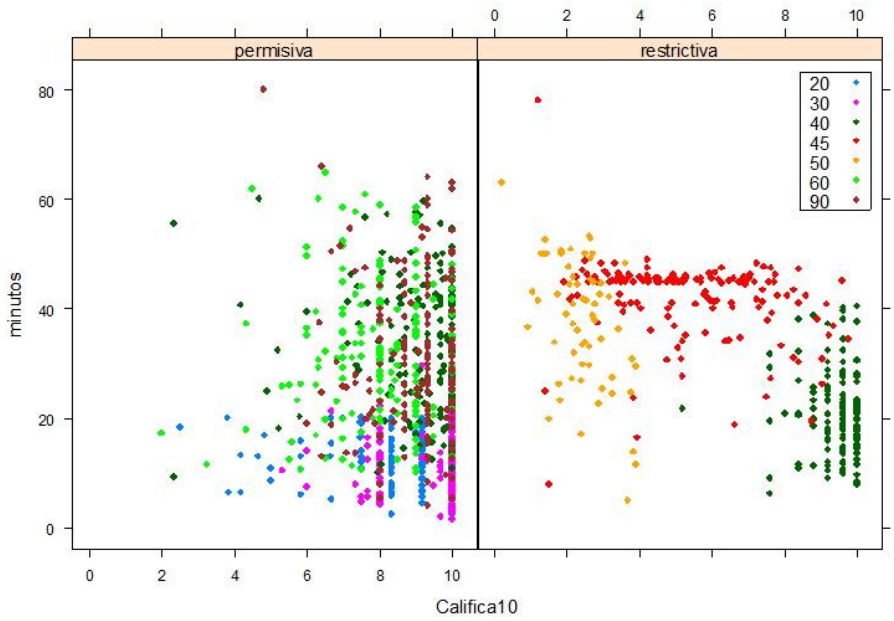


En este gráfico podemos ver los diferentes test según el tiempo que se les había dado para responder cada pregunta y vemos que sigue sin esclarecer el patrón en los permisivos, por lo que la teoría de la mala praxis del alumno tomaría más fuerza.

Si consideramos la variable respuesta ratioT, en vez de minutos, entonces el gráfico es el siguiente:



De este gráfico vemos como hay gente que con un % de tiempo ínfimo consigue el 10… frente a otros que utilizan el 100% del tiempo y no lo alcanzan.



En este gráfico se observa, dentro de las limitaciones por la nube de puntos, cómo no hay un patrón con la nota y lo que se tarda en responder, incluso recogemos algunos fallos de la plataforma indicando tiempo de realización superiores a los permitidos (seguramente problemas de sesiones mal cerradas).

En definitiva, por un lado tenemos la mala praxis del alumno en los test permisivos, y por otro lado se tiene que el docente da un tiempo de realización muy superior al permitido. Y esto puedo permitir que el alumno busque por diferentes fuentes las respuestas.

Remarcar como después de las malas calificaciones del test A3b2 y la nota media de la repesca similar a la del original, el A3b3 resulto más fácil que alguno de los permisivos vistos los valores obtenidos en los percentiles de dicho test.

4. CONCLUSIONES

Hemos comprobado que resulta muy fácil ver cuando un test (u otro tipo de prueba) ha sido demasiado fácil o difícil mediante una curva de frecuencias de las calificaciones obtenidas. Del mismo modo, mediante curvas de frecuencia o diagramas de caja, se puede ver si cada pregunta de un test ha resultado fácil o difícil de responder por el grupo.

Los tiempos para responder los cuestionarios son importantes, especialmente cuando el test se aplica en un entorno no controlado, como es el caso de la docencia a distancia. El fijar tiempos demasiado largos permite responder las preguntas buscando las respuestas en el material de lectura, y permite tomar nota de las propias preguntas para pasárselas a otros alumnos. El poner un tiempo corto dificulta estos tipos de mala praxis y mejora el test como herramienta de evaluación para medir el nivel de aprendizaje. De esta experiencia surge que un tiempo entre 1’30’’ y 2’00’’ por respuesta es suficiente, dependiendo también de la dificultad de las preguntas.

Si bien la retroalimentación inmediata es un mecanismo muy útil como reforzador del aprendizaje, la retroalimentación diferida en el tiempo es necesaria cuando queremos usar los cuestionarios como herramientas de evaluación del aprendizaje, más que como herramientas para reforzarlo. Lo ideal es que un test online se haga a una misma hora por todos los alumnos. Como esto no siempre es posible, se debe al menos evitar que el sistema muestre la retroalimentación sobre las respuestas correctas, al menos hasta el cierre definitivo del cuestionario.

Algunas plataformas de eLearning como Moodle, guardan información detallada de la realización de cada test, como la fecha y hora de realización, y el tiempo empleado. Esta información permite detectar ciertas anomalías: p. ej., si un alumno responde el test en un tiempo excesivamente corto y obtiene una calificación muy alta, es probable que ya tuviese en su poder las respuestas a las preguntas.

5. REFERENCIAS BIBLIOGRÁFICAS

Attwell G. (ed.), Evaluating E-learning: A Guide to the Evaluation of E-learning, Evaluate Europe Handbook Series, Vol. 2, ISSN 1610-0875, 2006

Govindasamy T., Successful implementation of e-Learning Pedagogical considerations, in *Internet and Higher Education*, vol. 4, pp. 287-299, 2002

Levine, S.J., Evaluation in Distance Education, in *Encouraging Learning:*

The Challenge of Teaching at a Distance, cap.2, 2003, LearnerAssociates.net http://www.LearnerAssociates.net

Rice Knowledge Bank, e-Learning for Development (curso online)

http://www.knowledgebank.irri.org/eLearningForDev/

Scheuermann F., Julius Björnsson (eds.), The Transition to Computer-Based Assessment: New Approaches to Skills Assessment and Implications for Large-scale Testing, European Commission, Joint Research Centre, ISBN 978-92-79-11110-5, DOI 10.2788/60083, 2009

Thurlow., M. et al., Computer-based Testing: Practices and Considerations, in Synthesis Report 78, National Center on Educational Outcomes, 2010

http://www.cehd.umn.edu/NCEO/onlinepubs/Synthesis78/Synthesis78.pdf

Twomey, E., Is there a role for computer-based assessment?, 1996. Accedido en agosto de 2007: http://science.universe.edu.au/mirror/CUBE96/twomey.html

---

[i] MCQ : "multiple choice questionnaire" en la literatura anglosajona.

[ii] http://www.knowledgebank.irri.org/eLearningForDev/

[iii] Por razones de protección de datos, y para evitar posibles agravios comparativos, preferimos no dar detalles específicos sobre el curso en cuestión, los profesores y los alumnos involucrados. La información utilizada se ha "anonimizado" para su tratamiento estadístico con fines de investigación.

[iv] Ver: http://www.r-project.org/