3690_Diseño y evaluación de herramientas para ayudar en la gestión de clases masificadas

Juan Ramón Rico-Juan; Antonio Javier Gallego; Jorge Calvo-Zaragoza; José Javier Valero-Más; David Rizo

juanramonrico@ua.es (Juan Ramón Rico-Juan), jgallego@dlsi.ua.es (Antonio Javier Gallego), jcalvo@dlsi.ua.es (Jorge Calvo-Zaragoza), jjvalero@dlsi.ua.es (José Javier Valero-Más), drizo@dlsi.ua.es (David Rizo)

Departamento de Lenguajes y Sistemas Informáticos, Universidad de Alicante, Carretera San Vicente del Raspeig s/n, Alicante, 03690, Spain

RESUMEN

Ante la necesidad de mantener la calidad en el proceso de enseñanza aprendizaje en grupos numerosos surgen nuevos retos para los profesores. Por un lado, si consideramos aceptable realizar un determinado número de actividades abiertas y creativas para evaluar a un grupo pequeño o mediano, el mantener este mismo número de actividades para un grupo numeroso hace que el esfuerzo del profesor se incremente desmesuradamente en la fase de evaluación, o bien, que se opte por disminuir el número de actividades o recurrir al sobre-explotado test de preguntas cerradas.

Este artículo propone un nueva metodología basada en corrección por pares (peer-review) y rúbricas (escala cerrada tipo Likert) que cumple un par de objetivos: aliviar la carga del docente en el proceso de evaluación e incrementar el aprendizaje del estudiante midiendo su grado de preparación en la revisión de actividades de compañeros. De este modo se mantiene la calidad en el proceso de enseñanza-aprendizaje con un número de actividades similar al de grupos pequeños o medianos.

Esta metodología ha sido implementada mediante varias herramientas de Google Drive y ha sido probada en una asignatura con dos grupos heterogéneos y con dos actividades distintas para medir la confianza y fiabilidad del sistema. Como resultado se han obtenido valores superiores al 95% en la confianza del test de correlación intraclase (ICC - Intraclass Correlation Coefficient) entre las calificaciones obtenidas por el sistema semiautomático y las del profesor. Estos datos confirman que el sistema propuesto tiene una respuesta similar a la del profesor sin que éste tenga que revisar exhaustivamente todas y cada una de las actividades entregadas por los alumnos.

Palabras clave: Revisión por pares; Rúbrica; Sistema inteligente; Evaluación

1. INTRODUCCIÓN

Hoy en día es común encontrar aulas en las que el número de estudiantes es alto. Esto implica no sólo una enseñanza menos enfocada sino también un inconveniente al proponer actividades a los estudiantes. Para aquellas actividades en las que la respuesta correcta es claramente objetiva, como las pruebas de elección múltiple o las que son numéricamente exactas -como pueden ocurrir en materias como la matemática o la física- el esfuerzo de corrección es menor y se puede hacer uso de algún tipo de herramienta de corrección automática. Sin embargo, hay una serie de temas en los que las actividades deben tener un alto grado de subjetividad y/o creatividad, por lo que la corrección debe ser realizada por un profesor. Dentro de este escenario, tradicionalmente hay dos opciones: (i) asumir toda la carga de corrección de la actividad, con la evidente aumento de carga de tiempo en grupos

masificados; (ii) optar por otros tipos de actividades que no requieran una evaluación tan amplia, lo que puede resultar poco conveniente en esos temas.

Una opción para paliar esta situación es recurrir a la revisión por pares, es decir, permitir que los estudiantes evalúen el trabajo de sus compañeros. Esto puede ser muy positivo desde el punto de vista de los alumnos, ya que la evaluación de otros trabajos les puede traer beneficios como una mayor comprensión de la tarea o proporcionarles otros puntos de vista sobre el mismo tema.

Sin embargo, es posible que los estudiantes no estén preparados para evaluar adecuadamente el trabajo. Por tanto, un proceso de revisión por pares requiere que el profesor verifique las correcciones hechas y que realice una estimación del desempeño de cada estudiante en relación a la calidad de su actividad. Por lo tanto, el profesor tiene que comprobar, no sólo el trabajo realizado por el estudiante, sino también las correcciones que ha recibido de sus compañeros, desembocando en un esfuerzo similar o incluso mayor que si se hubiera seguido el sistema de corrección tradicional.

Como resultado de esta red docente se presenta una metodología alternativa para implementar la revisión por pares y automatizar su corrección, lo cual aporta un beneficio directo para el alumno y además aligera la carga de trabajo de los docentes. Para ello, proponemos una metodología de corrección semi-supervisada en la que el sistema es capaz de detectar automáticamente si una corrección llevada a cabo por un alumno está sesgada con respecto a las correcciones de sus compañeros. Para verificar la bondad de esta metodología, se realizó un estudio con dos grupos de estudiantes totalmente heterogéneos en el que las correcciones realizadas por el sistema propuesto son comparadas con la corrección manual por parte del profesor. Los resultados obtenidos demuestran la alta precisión del sistema propuesto con una reducción significativa de la carga de trabajo con respecto a la corrección totalmente manual.

2. MÉTODO

2.1 REVISIÓN POR PARES

La revisión por pares o revisión por iguales (peer-review) ha sido ampliamente aplicada a lo largo de los años en el mundo científico. En este sistema las revisiones dadas por los científicos (iguales) se utilizan para decidir si un documento debe ser aceptado o rechazado (Hames, 2008). Sin embargo, la revisión por pares puede presentar importantes desventajas en algunos contextos en los que se podría poner en duda la objetividad (Wenneras y Wold, 1997). Además, algunos autores afirman que la revisión por pares sólo puede realizarse con éxito si los involucrados tienen una idea clara de su propósito fundamental (Horrobin, 1990).

2.2 METODOLOGÍA PROPUESTA

El uso del sistema de revisión por pares para la corrección de actividades en clases masificadas está justificado por varios motivos. En primer lugar para aliviar la carga de trabajo del profesor, el cual, gracias a este tipo de corrección puede permitirse plantear todas las actividades necesarias para la formación de los alumnos en la materia a tratar; de otro modo debería de considerar la posterior carga de trabajo y consecuentemente limitar la cantidad y tipo de actividades a realizar. Además, este proceso permite a los estudiantes asimilar mejor los conceptos relacionados con las actividades, ya que también tienen que actuar como evaluadores, y dado que la revisión por pares es parte de la actividad, también sirve para medir el grado de toma de decisiones del estudiante. Es decir, cuando un evaluador indica que la respuesta a una pregunta es correcta o no, está demostrando la comprensión del concepto evaluado, y por lo tanto las evaluaciones correctas conllevarán mejores calificaciones.

Para que la evaluación final de la actividad sea lo más justa y consensuada posible es nece-

sario realizar varias evaluaciones sobre cada una de las actividades. De esta manera, el cálculo de la mediana de las decisiones permitirá establecer una puntuación cercana a la que asignaría el docente. Aún así, el profesor debe revisar alguna de estas evaluaciones para asegurar que la puntuación final corresponda a la calidad de la actividad. A diferencia de otros escenarios, como la de los artículos científicos, en docencia podemos asumir que existe una evaluación perfecta para una determinada actividad, que es la proporcionada por el profesor. Esto, por un lado nos permitirá validar o revisar las evaluaciones de los alumnos en las que detectemos que no se ha llegado a una corrección consensuada, y por otro también nos permitirá estudiar la fiabilidad y confianza de los sistemas propuestos para automatizar la corrección.

Al principio puede parecer que este escenario incluso aumenta la carga de trabajo del profesor. Es por eso que la idea principal es asumir que aquellas evaluaciones que tienen una puntuación relativamente consensuada o sin casos atípicos es porque las correcciones son justas. Por lo tanto, sólo las correcciones atípicas con respecto al resto deberán ser revisadas de nuevo por el profesor.

En la Figura 1 podemos ver el esquema que hemos planteado para abordar el problema, el cual se divide en los siguientes tres pasos:

1) Entrega de actividades y asignación balanceada de revisiones a los alumnos.
2) La revisión por pares por parte de los alumnos.
3) La supervisión del profesor de todo el proceso y la revisión de las evaluaciones en las que no se llegue a consenso o se detecten valores atípicos.

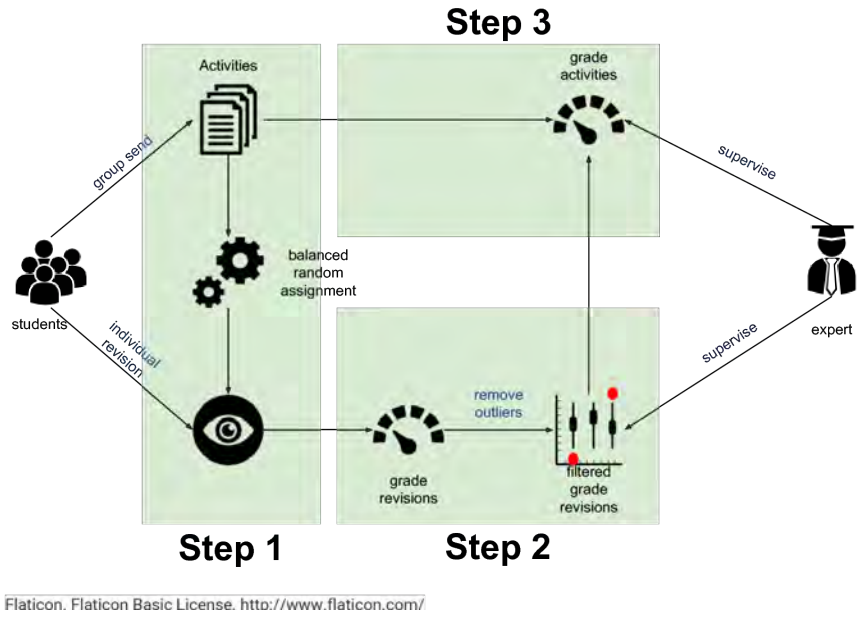


Figura 1: Esquema de la metodología aplicada.

En el paso 1 del proceso se recopilan todas las entregas de los alumnos y se realiza el sorteo de las actividades que tendrá que corregir cada alumno (obviamente excluyendo la propia). A cada alumno le corresponderá revisar varias actividades, cuantas más revisen (lo cual dependerá de la longitud de la revisión) nos permitirá recabar un mayor número de correcciones por actividad y por lo tanto calcular la nota con un mayor nivel de seguridad. En el paso 2 cada alumno tendrá que revisar las actividades que se le hayan asignado y enviar sus correcciones mediante el medio estipulado (en nuestro caso un formulario de Google Forms). Una vez obtenidas todas las evaluaciones se calculará

la nota a proponer para cada actividad utilizando la metodología descrita y se analizarán las actividades para las que no se ha llegado a un consenso o en las que han aparecido caso atípicos (correcciones muy alejadas de la considerada revisión correcta). Para estos dos casos se avisará al profesor (paso 3), el cual tendrá que revisar únicamente estas actividades, asignando su evaluación como nota final del trabajo.

Pero, ¿cómo calculamos la nota a partir de las evaluaciones de los alumnos? En la siguiente sección explicamos el método propuesto mediante un ejemplo sencillo.

2.2.1 Ejemplo sencillo e ilustrativo sobre la metodología propuesta

Consideremos una actividad y 3 revisores: Alex, Jane y Víctor. La revisión consiste en la evaluación de 3 preguntas o ítems utilizando una escala de valoración de 0 a 3, siendo 0 la peor puntuación y 3 la mejor.

Supongamos las siguientes revisiones:

	ítem 1	ítem 2	ítem 3
Alex	1	1	1
Jane	1	1	3
Víctor	3	3	2

Con estas respuestas, ¿qué puntuación asignamos a cada uno de los ítems? Para esto podemos seguir distintos criterios. La plataforma Moodle cuenta con un complemento llamado Workshop que evalúa la revisión por pares. Para este complemento la mejor revisión es aquella que está más centrada geométricamente. Es equivalente a la definición de la **mediana de un conjunto** de elementos con varias dimensiones. Es decir, alguno de los revisores será el mejor y será al que se le hará caso exclusivamente para poner la nota.

Criterio de la mediana de un conjunto

En este criterio en primer lugar calculamos las distancias entre las evaluaciones usando la distancia L1 (Manhattan) o la L2 (Euclídea), como se muestra en las siguientes tablas:

L1	Alex	Jane	Victor
Alex	0,0	2,0	5,0
Jane	2,0	0,0	5,0
Víctor	5,0	5,0	0,0

L2	Alex	Jane	Victor
Alex	0,0	2,0	3,0
Jane	2,0	0,0	3,0
Víctor	3,0	3,0	0,0

A partir de estos datos ya podemos calcular cual sería la “mejor revisión”, que sería aquella cuya suma de distancias por fila sea menor. Es decir, siendo la variable *i* el revisor y *j* el ítem o pregunta evaluada, definiríamos la mejor revisión como:

En la siguiente tabla se puede ver el cálculo de la mejor revisión en base a las distancias L1 y L2 obtenidas anteriormente.

Suma de distancias	L1	L2
Alex	7,0	5,0
Jane	7,0	5,0
Víctor	10,0	6,0

Según este criterio la mejor revisión le corresponde a Alex con sus evaluaciones (1, 1, 1) o a Jane

con (1, 1, 3). Como comprobamos en este caso aparece una ambigüedad para seleccionar cuál es la mejor revisión con sus correspondientes ítems, ya que existen dos posibilidades.

Además, si sólo utilizamos la valoración de un evaluador es posible que, aunque su evaluación tenga un valor o criterio medio con respecto a los demás, para algunas preguntas haya corregido de forma incorrecta, lo cual sería injusto para el alumno evaluado. Por lo que, ¿no sería mejor aprovechar todas las revisiones para tener un criterio más consensuado?

Criterio de la mediana por ítem (propuesta para seleccionar la mejor evaluación)

Nuestra propuesta es utilizar el concepto de **mediana por ítem**. Este concepto representa una evaluación más centrada que la definición anterior de mediana de un conjunto. Para esto tenemos que calcular el resultado medio de cada elemento a partir de la matriz de revisiones, la cual la trataremos como una matriz multidimensional de números reales. A continuación se puede ver el cálculo de la mediana para las evaluaciones del ejemplo anterior.

	ítem 1	ítem 2	ítem 3
Alex	1,0	1,0	1,0
Jane	1,0	1,0	3,0
Víctor	3,0	3,0	2,0
	-----	-----	-----
Mejor revisión (mediana)	1,0	1,0	2,0

En este caso, la mejor revisión sería (1, 1, 2). Utilizando esta media podemos calcular la distancia o discrepancia para cada revisor, lo que nos indicará cómo de cercana o similar es su corrección con respecto a la mejor revisión.

Distancia de las revisiones a la mejor revisión

Una vez obtenida la mejor revisión procedemos a calcular la distancia desde ésta hasta las revisiones realizadas por los evaluadores. En la siguiente tabla se puede ver el desglose del cálculo de la distancia L1 entre la mejor revisión y cada una de las correcciones, añadiendo una columna final con el sumatorio de las distancias en valor absoluto.

L1(i, mejor_revisión)	ítem 1	ítem 2	ítem 3	**|Suma|**
Alex	0,0	0,0	-1,0	1,0
Jane	0,0	0,0	1,0	1,0
Víctor	2,0	2,0	0,0	4,0

Siguiendo con el ejemplo, procedemos a realizar estos cálculos tanto para la distancia L1 como para la L2, obteniendo finalmente los siguientes resultados para el sumatoria de las distancias por evaluador:

distancia(i, mejor_revisión)	**L1**	**L2**
Alex	1,0	1,0
Jane	1,0	1,0
Víctor	4,0	2,8

Como se puede apreciar, las discrepancias de la revisión individual se reflejan directamente cuando usamos la distancia L1 por lo que es ésta la que usaremos en nuestra metodología.

A partir de estos cálculos procedemos a determinar el umbral utilizado para discriminar las evaluaciones atípicas. Para esto analizaremos los cuartiles de la distribución formada por las distancias calculadas. Si por ejemplo obtenemos una distribución con Q3=2 (donde Q3 se refiere al 3er cuartil de la distribución) y con un IQR=1 (donde IQR es el *Inter Quantile Rate* o rango intercuartílico, que es la diferencia entre el tercer y el primer cuartil de una distribución), podemos fijar el umbral (U) de los valores atípicos en:

$$U = Q3 + 1.5 * IQR = 2 + 1.5 * 1 = \mathbf{3.5}$$

Por lo tanto, Víctor, que tenía 4 discrepancias entre su corrección y la mejor corrección (ver la distancia L1 de la tabla anterior), superaría este umbral, por lo que se le consideraría una revisión atípica. En este caso su revisión sería descartada y se volverían a calcular la mediana para obtener la mejor revisión pero sin considerar la de Víctor.

Si analizamos las puntuaciones de Víctor podemos ver cómo evaluó los ítems 1 y 2 con una desviación de 2 puntos con respecto al valor de consenso. La acumulación de discrepancias hace que finalmente sea considerada como una evaluación atípica, aunque el ítem 3 fuese correctamente evaluado.

De esta forma podríamos obtener una evaluación final a partir de las correcciones que sí que han llegado a consenso, además, en el caso de las evaluaciones con varias correcciones atípicas podríamos avisar al docente para que las revise manualmente. Es decir, nuestra propuesta consistiría en un sistema **semiautomático** que en general devolvería la calificación final sin intervención humana, pero que en determinados casos preguntaría al experto. Esta característica además nos permitirá determinar cuán buenos evaluadores son nuestros alumnos, pudiendo otorgarles una nota por ello.

Asignando una calificación a cada actividad

Para el cálculo de la calificación final de cada actividad, una vez descartadas las revisiones atípicas, simplemente tendríamos que calcular la mediana de las revisiones que sí que han llegado a consenso. Además, si lo consideramos oportuno, podemos ponderar el peso de cada ítem en la calificación final. A continuación se incluye un ejemplo de este cálculo.

Evaluación	ítem 1	ítem 2	ítem 3	
Pesos	30%	30%	40%	Sumatorio de los pesos = 100%
mejor revisión (mediana)	1,0	1,0	2,0	Calificación final = (30*1+30*1+40*2)*10/(3*100) = 4.67

El resultado de la calificación final sería de un 4.67 sobre 10. Para obtener este resultado primero hemos calculado la mediana de las mejores revisiones (descartando las atípicas), a continuación hemos ponderado el peso de cada ítem (asignando un peso sobre 100% para cada pregunta), obteniendo (30*1+30*1+40*2)/100. Y además hemos normalizado el resultado de la nota entre 0 y 10, ya que la escala de la rúbrica (como ya vimos antes) estaba establecida entre 0 y 3, por lo que además multiplicamos por 10 y dividimos entre 3, obteniendo que:

$$\text{Calificación final} = (30*1+30*1+40*2)*10/(3*100) = 4.67$$

Mediante este proceso podemos obtener de forma automatizada la calificación para todas las acti-

vidades, interviniendo únicamente en los casos a los que no se haya llegado a consenso. A continuación vamos a analizar los resultados obtenidos de la experimentación en un caso real.

3. RESULTADOS

El objetivo del experimento es aplicar nuestro método a diferentes escenarios para analizar su confianza y fiabilidad. Por este motivo se realizó un estudio durante el curso académico 2015-16, en el que se aplicó la metodología propuesta a dos grupos disjuntos de estudiantes para la evaluación de dos actividades diferentes: la primera sobre propiedad intelectual y la segunda sobre la Webquest. Se establecieron entregas de actividades por grupos (formados por 2 personas en la mayoría de los casos), y para el proceso de evaluación se asignaron 5 revisiones por alumno, por lo que obtuvimos una media de 10 revisiones diferentes por actividad.

Además, para tratar de asegurar que las revisiones fueran imparciales, parte de la nota del estudiante (30%) dependía de si sus revisiones eran correctas o no (las consideradas como revisiones atípicas). El resto de la nota (70%) correspondía a la calificación de la propia actividad. Estos porcentajes se estimaron en base al tiempo que el estudiante dedicó a cada parte.

Todo el proceso de revisión fue asistido por un experto y se realizó en base a unas rúbricas para facilitar el proceso.

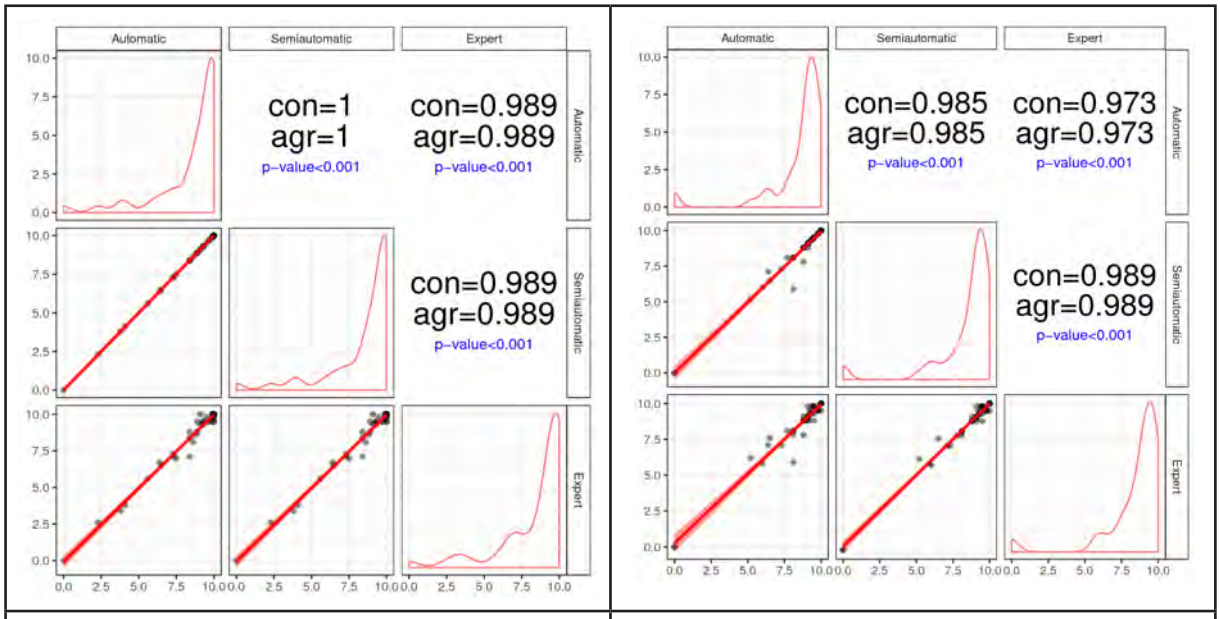
Para contrastar los resultados obtenidos se van a analizar tres escenarios diferentes según el proceso de evaluación:

1) **Automático**: el sistema propone directamente las calificaciones finales de las actividades y descarta las correcciones atípicas, pero no consulta al experto para su corrección.
2) **Semiautomático**: el sistema propone automáticamente las calificaciones para todas las actividades en las que se ha llegado a consenso en la evaluación, y pregunta al profesor experto sólo para las que no haya consenso.
3) **Experto**: todas las actividades son evaluadas manualmente por un profesor experto.

3.1 Primera actividad: propiedad intelectual

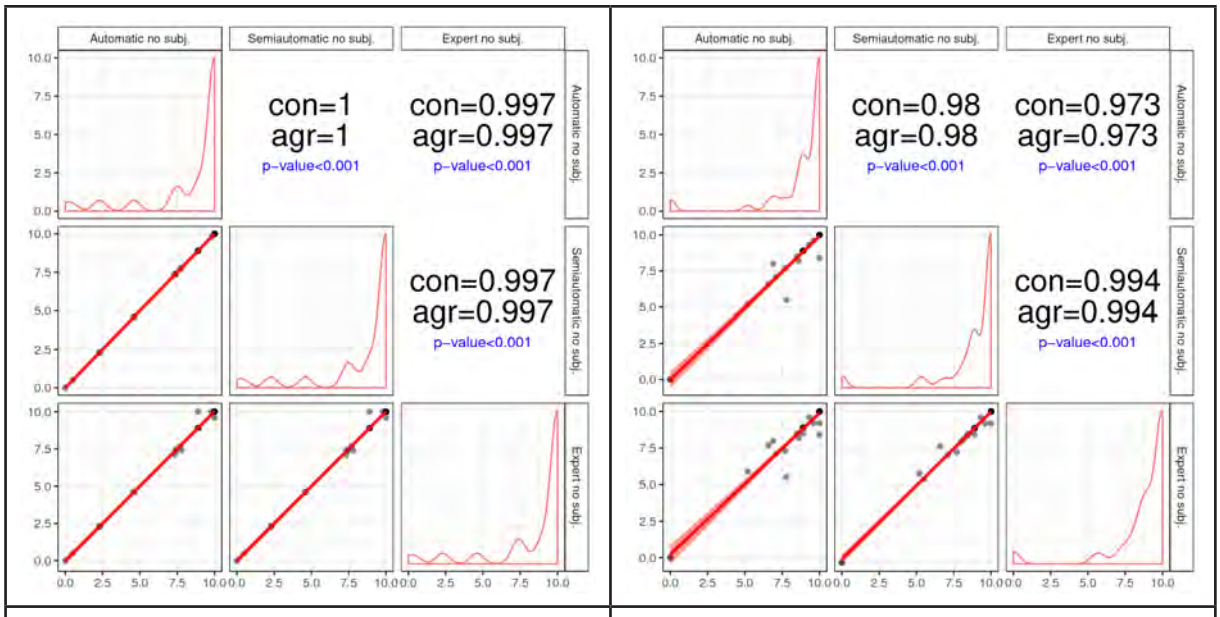
En esta sección vamos a analizar los resultados obtenidos para la primera actividad considerando los tres escenarios establecidos: evaluación automática, semiautomática y la evaluación de un experto. Además hemos diferenciado entre las preguntas que son puramente objetivas y las que tenían un cierto grado de subjetividad, lo que nos permitirá validar nuestra propuesta diferenciando también estos casos.

Con el fin de probar la fiabilidad de las medidas pareadas con respecto a la misma evaluación (Watson y Petrie, 2010), aplicamos el coeficiente de correlación intraclase (ICC) para este propósito (Bartko, 1966). En la Figura 2 se pueden ver los resultados obtenidos para los dos grupos considerando todas las respuestas (objetivas y subjetivas), y en la Figura 3 cuando consideramos sólo las respuestas objetivas. Observamos la alta consistencia y concordancia obtenida en la comparación de calificaciones Automático vs. Experto, cuyos valores son superiores a 0,97 (sobre 1), y superior a 0,98 en el caso Semiautomático vs. Experto (nuestra metodología propuesta).



Grupo 1 | Grupo 2

Figura 2: Comparación de la actividad sobre la propiedad intelectual incluyendo todas las respuestas en diferentes escenarios. La diagonal (empezando en la esquina superior izquierda hasta la esquina inferior derecha) muestra gráficas sobre la densidad de las observaciones, mientras que la matriz triangular inferior y superior muestran gráficas y resultados numéricos con los coeficientes de correlación intraclase: consistencia (con.) y concordancia (agr.), respectivamente.



Grupo 1 | Grupo 2

Figura 3: Comparación de la actividad sobre la propiedad intelectual usando únicamente las respuestas objetivas en los diferentes escenarios. La diagonal (empezando en la esquina superior izquierda hasta la esquina inferior derecha) muestra gráficas sobre la densidad de las observaciones, mientras que la matriz triangular inferior y superior muestra gráficas y resultados numéricos con los coeficientes de correlación intraclase: consistencia (con.) y concordancia (agr.), respectivamente.

Como se puede ver en los resultados anteriores, en el caso de utilizar respuestas concisas (sin sub-

jetividad), este coeficiente apenas aumenta respecto del de las respuestas completas.

Con el fin de observar el grado de diferencias entre los distintos sistemas, podemos ver en la Figura 4 como en los resultados del experto en comparación con los del semiautomático (que sería nuestra propuesta), y tanto considerando como sin considerar las respuestas subjetivas, aparece una mayor densidad alrededor de cero (que sería el mejor valor). Sin embargo, se puede observar que en el caso del experto vs. automático, los valores también se concentran alrededor del mismo valor. Esto quiere decir que podríamos optar por una aproximación completamente automática en este caso, aunque al tratarse de la puntuación de trabajos de alumnos es preferible que cuando haya discrepancias tenga que ser revisado por un experto.

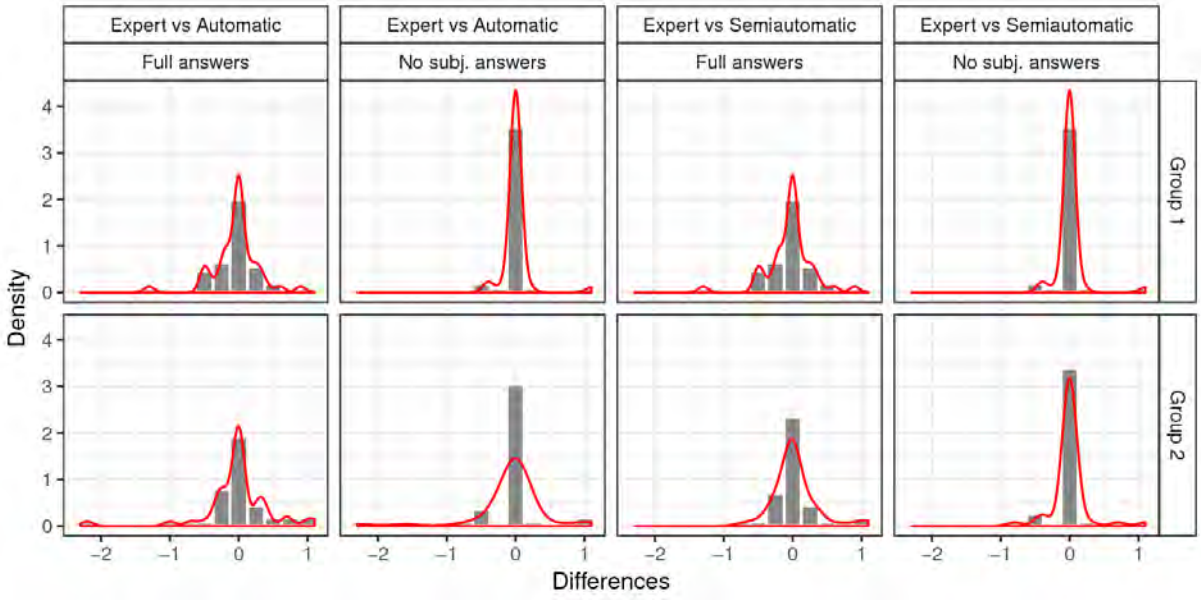


Figura 4: Comparativa de la actividad sobre la propiedad intelectual respecto a las diferencias entre calificaciones finales obtenidas desde los diferentes puntos de vistas abordados en la experimentación según grupos, tipos de preguntas (completas o solo objetivas) y las propuestas de evaluación (Experto frente a sistema automático o semiautomático).

3.2 Segunda actividad: Webquest

En este segundo apartado vamos a realizar el mismo análisis para la segunda actividad planteada, la realización de una Webquest. En este caso también evaluaremos de forma automática, semiautomática y por un experto, y consideraremos las respuestas completas (objetivas y subjetivas) (ver Figura 5) o solo las respuestas objetivas (Figura 6).

Al igual que en la actividad anterior, vamos a aplicar el coeficiente de correlación intraclase. En este caso se mantiene también una alta consistencia y concordancia entre los diferentes sistemas manteniendo una ligera diferencia entre el grupo 1 y el grupo 2, concretamente de 0.99 en el primer caso y del 0.94 en el segundo (Experto vs. Semiautomático), como se puede ver en la Figura 5. Seguramente esto se debe a la complejidad y variabilidad de esta actividad, y a la diferencia general entre los propios estudiantes de los grupos. El segundo grupo presentaba un rendimiento claramente inferior que el primero. En la Figura 6, donde se estudia el comportamiento de la calificaciones atendiendo a la respuestas objetivas, obtenemos unos coeficientes similares a los de las respuestas completas, siendo los valores del grupo 2 ligeramente superiores.

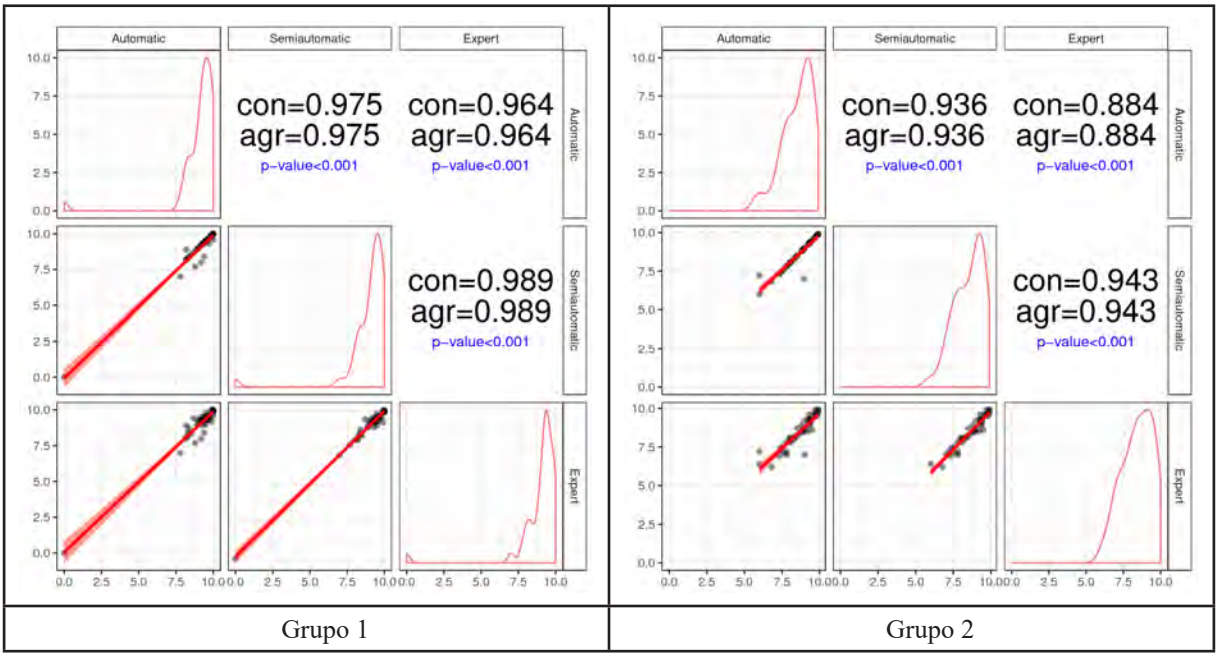


Figura 5: Comparación de la actividad sobre Webquest con respuestas completas en los diferentes escenarios. La diagonal (empezando en la esquina superior izquierda hasta la esquina inferior derecha) muestra gráficas sobre la densidad de las observaciones, mientras que la matriz triangular inferior y superior muestra gráficas y resultados numéricos con los coeficientes de correlación intraclase: consistencia (con.) y concordancia (agr.), respectivamente.

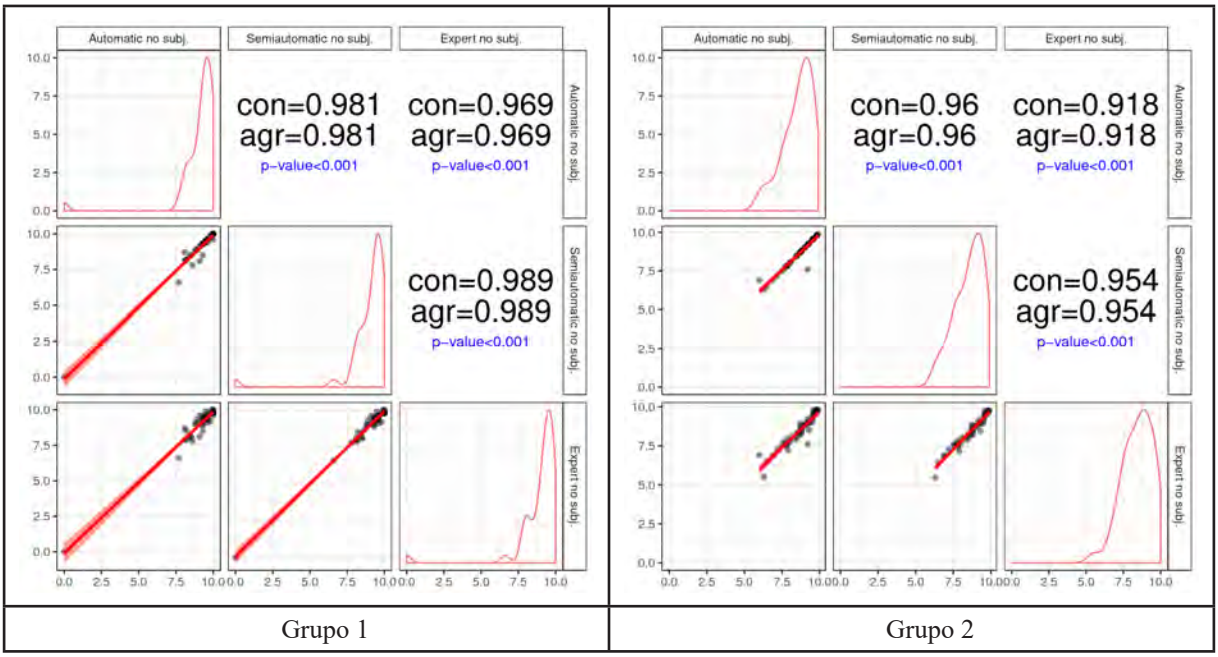


Figura 6: Comparación de la actividad sobre Webquest usando únicamente las respuestas objetivas en diferentes escenarios. La diagonal (empezando en la esquina superior izquierda hasta la esquina inferior derecha) muestra gráficas sobre la densidad de las observaciones, mientras que la matriz triangular inferior y superior muestra gráficas y resultados numéricos con los coeficientes de correlación intraclase: consistencia (con.) y concordancia (agr.), respectivamente.

Observando el grado de diferencias entre el sistema de evaluación de expertos contra el automático

o el semiautomático, podemos ver en la Figura 7 como Experto vs. Semiautomático y tanto para las respuestas no subjetivas como para las subjetivas, se alcanza la mayoría de los valores de densidad alrededor de cero (que se corresponde con el mejor valor). Los resultados son similares a los obtenidos en la actividad anterior.

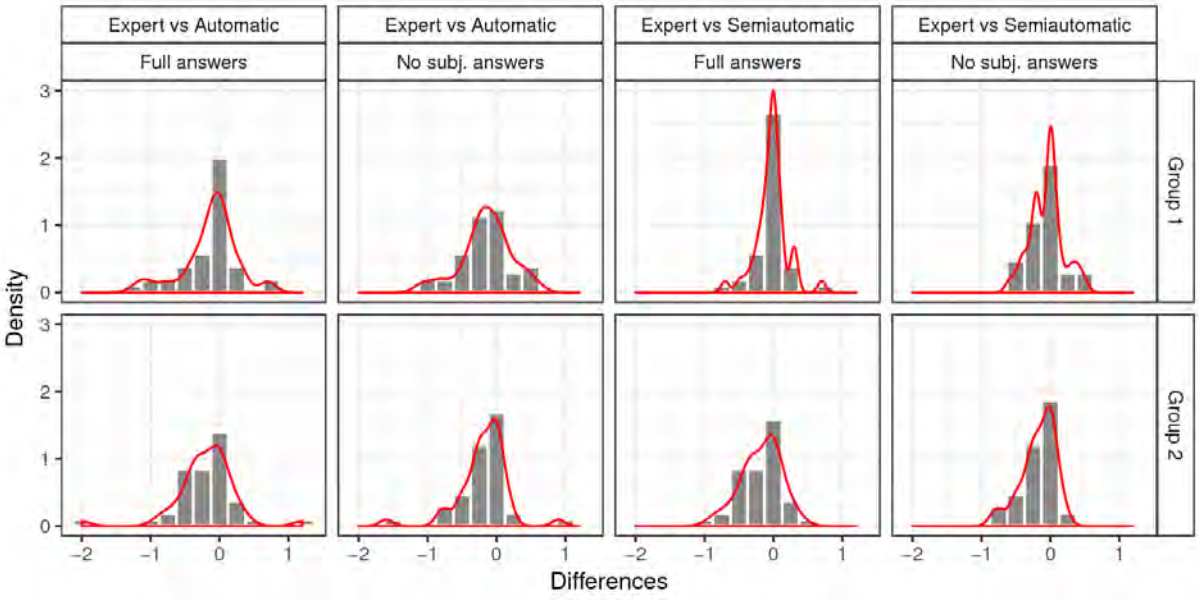


Figura 7: Comparativa para la actividad sobre la Webquest respecto a las diferencias entre calificaciones finales obtenidas desde los diferentes puntos de vistas abordados en la experimentación según grupos, tipos de preguntas (completas o solo objetivas) y propuestas de evaluación (Experto frente a sistema automático o semiautomático).

4. CONCLUSIONES

Hemos propuesto un nuevo sistema de corrección semi-supervisado para aliviar la carga de trabajo de los profesores en la evaluación de las actividades complejas donde interviene cierto grado de subjetividad. Nuestro enfoque se basa en ideas de revisión por pares, rúbricas y detección de casos atípicos. Para llevar a la prácticas este sistema se han utilizado los servicios básicos de Google Forms, Spreadsheets y Apps Script, a los cuales puede acceder por cualquier usuario de forma gratuita, no requiere instalación de ningún programa ni utilización de servidores propios, y además permiten compartir fácilmente esta solución o los resultados de los mismos simplemente realizando una copia de la hoja de cálculo.

Además de las ventajas mencionadas anteriormente, este sistema permite al profesor dedicar su tiempo a supervisar el proceso, únicamente ayudando a los estudiantes mientras realizan la revisión por pares durante las sesiones asistidas, y a la vez reduce considerablemente el tiempo dedicado a la corrección cuando consideramos un gran número de estudiantes y mantiene la calidad respecto al número de actividades como si se tratase de un grupo pequeño o mediano.

Para validar el método propuesto se ha realizado un estudio aplicándolo a la evaluación de dos prácticas realizadas por dos grupos de alumnos durante un curso académico. Las medidas intraclase superiores al 0.94 entre el sistema semiautomático y el sistema experto (profesor) validan el método propuesto, y su aplicación a diferentes problemas y grupos confirman su fiabilidad.

La primera propuesta para trabajos futuros consistiría en usar la opinión de varios expertos para cada actividad y estimar las discrepancias que existan entre ellos para compararlas con las generadas directamente por el sistema automático y semiautomático. Como segunda propuesta se podrían estudiar otros tipos de métodos basados en aprendizaje automático para detectar los casos atípicos y evaluar su impacto. Por último, se podría extender el uso de este nuevo sistema en el tipo de cursos MOOC con fechas de inicio y fin establecidas cuyo objetivo fuera obtener una calificación final automática o semiautomática.

6. TAREAS DESARROLLADAS EN LA RED

PARTICIPANTE DE LA RED	TAREAS QUE DESARROLLA
Juan Ramón Rico-Juan	Coordinación, implementación y aplicación de la metodología. Aportación de ideas base.
Antonio Javier Gallego	Implementación y aplicación de la metodología. Aportación de ideas base.
Jorge Calvo-Zaragoza	Aportación de ideas a la metodología.
José Javier Valero-Más	Aportación de ideas a la metodología.
David Rizo	Aportación de ideas a la metodología.

7. REFERENCIAS BIBLIOGRÁFICAS

Bartko, J. J. (1966). *The intraclass correlation coefficient as a measure of reliability*. Psychological reports, 19 , 3–11.

Hames, I. (2008). *Peer review and manuscript management in scientific journals: guidelines for good practice*. John Wiley & Sons.

Horrobin, D. F. (1990). *The philosophical basis of peer review and the suppression of innovation.* Jama, 263 , 1438–1441.

Watson, P., & Petrie, A. (2010). *Method agreement analysis: a review of correct methodology*. Theriogenology, 73 , 1167–1179.

Wenneras, C., & Wold, A. (1997). *Nepotism and sexism in peer-review*. Nature, 387 , 341.