

# Evaluando la auto-evaluación: relación entre la percepción del propio conocimiento y el ajeno

## Assessing self-assessment: self-perception of competency related to peer evaluation

Pilar Garcia-Almirall<sup>1</sup>, Ernest Redondo<sup>1</sup>, Francesc Valls<sup>1</sup>, David Fonseca<sup>2</sup>  
{pilar.garcia-almirall, ernesto.redondo, francesc.valls}@upc.edu, fonsi@salle.url.edu

<sup>1</sup> Escuela Técnica Superior de Arquitectura de Barcelona  
UPC - BarcelonaTech  
Barcelona, España

<sup>2</sup> La Salle, Campus Barcelona  
Universitat Ramon Llull (URL)  
Barcelona, España

**Resumen-** En procesos de aprendizaje, la adecuada percepción de las propias capacidades es imprescindible para ser capaz de detectar las carencias y para auto-dirigir el aprendizaje. En la presente comunicación se describe una experiencia educativa de evaluación entre iguales en el ámbito universitario, donde los estudiantes evaluaron su propio trabajo y dos trabajos ajenos seleccionados aleatoriamente. En ambos casos, la evaluación de los trabajos se hizo utilizando dos metodologías: mediante una rúbrica y con evaluación subjetiva. Los resultados muestran que no existe relación significativa entre la auto-evaluación y las evaluaciones recibidas, mientras que sí se observa entre esta y las evaluaciones realizadas, en ambos casos independientemente de la utilización o no de una rúbrica. La hipótesis de los autores es que en las fases iniciales de aprendizaje, el alumnado carece de las herramientas para evaluar otros trabajos y proyecta la percepción de su propio grado de conocimiento hacia el resto del estudiantado.

**Palabras clave:** *Evaluación entre iguales, Auto-evaluación, Educación, Auto-percepción, Efecto Dunning-Kruger, Arquitectura, Sistema de Información Geográfica*

**Abstract-** In learning processes, the adequate perception of one's capacity is crucial to detect one's shortcomings and to self-direct learning. This paper describes an educational peer-review experience in higher education, where students were asked to assess their own assignments as well as two randomly selected assignments. The evaluation of the assignments was performed using two methodologies: using a rubric and through subjective evaluation. The results show that while there is not a significant relationship between self-assessment and peer-assessments, a relationship is observed between self-assessment and the grades given to other participants, in both cases independently of whether a rubric was used or not. The authors' hypothesis is that in the initial phases of learning, students lack the tools to assess other assignments and project the perception of their own competency onto their peers.

**Keywords:** *Peer-assessment, Self-assessment, Education, Self-perception, Dunning-Kruger effect, Architecture, Geographic Information System*

### 1. INTRODUCCIÓN

En los procesos de aprendizaje en estudios avanzados (de grado o de post-grado) no es solamente necesario que el estudiantado adquiera los conocimientos requeridos, sino que además le deben ser proporcionadas las herramientas necesarias para poder auto-evaluarse adecuadamente, capacidad que le será necesaria para su adaptabilidad en un contexto de aprendizaje permanente.

El conocido como efecto Dunning-Kruger (Kruger y Dunning, 1999; Ehrlinger y Dunning, 2003), describe la percepción exageradamente alta de las propias capacidades en las fases iniciales de adquisición de conocimientos en una materia, así como la percepción más baja de lo que correspondería por parte los expertos en ella. Este efecto tiene implicaciones sobre el proceso educativo, puesto que en el primer caso la capacidad del alumnado de evaluar su propia competencia queda mermada por la falta de herramientas (experiencia o conocimientos) necesarias para ello, mientras que los instructores (como expertos) pueden juzgar las tareas que encomiendan a los estudiantes como más asequibles de lo que realmente son.

Por otro lado, en cuanto a la evaluación del conocimiento de otras personas en relación a la percepción del propio conocimiento, algunos experimentos (Ames y Kammrath, 2004) demuestran que una mayor percepción de propia competencia redundaba en una más baja evaluación de la competencia de otras personas, explicando este fenómeno a través del narcisismo de los sujetos.

Para investigar la relación entre la propia evaluación y (a) las evaluaciones recibidas o (b) las evaluaciones efectuadas, se realizó una experiencia educativa de evaluación entre iguales en el ámbito universitario –los resultados de la cual se exponen en esta ponencia–, donde se pidió al estudiantado que evaluara su propio trabajo y dos trabajos ajenos seleccionados aleatoriamente dentro de la misma clase, tanto mediante el uso de una rúbrica como con una evaluación subjetiva.

## 2. CONTEXTO

La evaluación entre iguales puede ser utilizada como herramienta educativa en dos aspectos fundamentales: (a) mediante la autoevaluación, fomentando que el alumnado efectúe una reflexión crítica sobre su trabajo, más allá de la corrección del profesorado; y (b) a través de la evaluación de trabajos de compañeros o compañeras, permitiendo conocer distintas estrategias para resolver un mismo problema.

Adicionalmente, la posibilidad de que el alumnado se autoevalúe no solo tiene repercusiones en su capacidad de aprendizaje autónomo, sino que tiene el potencial de reducir la carga lectiva del profesorado empleada en evaluar los trabajos, tiempo que puede ser empleado en una mejorar la docencia. En este sentido los *Massive Open Online Courses* (MOOCs), por las dificultades logísticas de su gran número de estudiantes, han utilizado la evaluación entre iguales (*peer-review*), así como otras estrategias para reducir la intervención del profesorado, como la evaluación automática (*robo-grading*) o el soporte de profesores voluntarios (*community TAs*).

El objetivo de la experiencia fue estudiar la posibilidad de utilizar la evaluación entre iguales en la enseñanza universitaria, específicamente en el campo de la arquitectura y en la enseñanza de Sistemas de Información Geográfica (SIG), para mejorar la adquisición de conocimientos y la experiencia de usuario del alumnado. Los autores creen que la experiencia es trasladable a otros campos de la educación superior y secundaria, así como a otros ámbitos de conocimiento.

## 3. DESCRIPCIÓN

La experiencia educativa se desarrolló en la asignatura optativa “Estudios urbanos con tecnología informática SIG”, que se desarrolló en el curso académico 2014-2015 (impartida del 21 al 27 de enero de 2015) en la Escuela Técnica Superior de Arquitectura de Barcelona. La asignatura es de tipo intensiva, desarrollándose en aproximadamente una semana, y consta de un total de 4.5 créditos ECTS (European Credit Transfer and Accumulation System). La enseñanza de esta asignatura (García-Almirall, Redondo, Valls, y Corso, 2014), desarrollada durante casi 20 años, es eminentemente práctica, utilizando prácticas guiadas con una progresiva autonomía del alumnado para resolver los problemas que se plantean a lo largo de curso (*worked examples with faded guidance*). Estos ejercicios se intercalan con clases teóricas en las que se introducen los temas que se trabajarán en los talleres, de manera que ambos tipos de enseñanza se complementan, concluyendo con un ejercicio autónomo (desarrollado en horas no lectivas) donde el estudiante desarrolla una propuesta de innovación utilizando las herramientas tecnológicas que ha aprendido, que se presenta y discute en grupo el último día del curso.

La idea de lugar es central en la enseñanza de arquitectura; en ediciones anteriores, las prácticas del curso se centraban en el estudio de caso de una zona urbana concreta, evitando así introducir múltiples contextos de trabajo, y permitiendo a la vez el análisis de un mismo lugar con criterios progresivamente más complejos. Para la experiencia de evaluación entre iguales, se planteó una división asimétrica de las prácticas en dos partes, siendo la primera de ellas la que fue objeto del experimento. Esta primera parte se desarrolló en un ámbito de escala regional, con el doble objetivo de practicar el uso de los SIG a una escala diferente a la urbana, y al mismo tiempo permitir al alumnado conocer su avance (a través de la entrega de la práctica y la

posterior evaluación de sus compañeros y compañeras) en las primeras fases del desarrollo del curso.

El proceso seguido para la evaluación fue el siguiente (García-Almirall, Redondo, Valls, & Fonseca, 2015): (1) el alumnado entregó sus trabajos en la intranet docente (basada en Moodle) y los documentos fueron anonimizados, eliminando cualquier referencia a los autores tanto en el contenido como en los metadatos; (2) se asignó un número a cada trabajo y a continuación se emparejaron aleatoriamente de manera que a cada participante le eran asignados dos trabajos para ser evaluados; y (3) se utilizó la plataforma Google Forms (Google, 2015) para recoger las evaluaciones tanto de los dos trabajos asignados como del trabajo propio.

La evaluación de los trabajos se estructuró en dos partes, en la primera utilizó una rúbrica para valorar seis aspectos distintos (compleción, presentación, estructura, salidas gráficas, conclusiones, y redactado) en una escala de 0 a 3 puntos (insatisfactorio, cumple mínimos, satisfactorio, o excepcional) de manera que la evaluación máxima posible era de 18 puntos (que se convirtieron posteriormente a una escala de 0 a 10). En la segunda parte, se evaluaron dos aspectos generales sin el uso de una rúbrica, en una escala de 0 a 10 (calidad general del trabajo y nivel de comprensión adquirido).

## 4. RESULTADOS

El grupo matriculado constaba de 18 personas (de las cuales 5 eran mujeres y 13 hombres), siendo finalmente 16 personas quienes formaron parte del experimento. De ellos, cinco participantes olvidaron realizar su auto-evaluación de manera que tuvieron que ser descartados para los análisis, quedando reducido a 11 el número de muestras del experimento. Los datos se descargaron de la hoja de cálculo de Google Drive vinculada al formulario de Google Forms y se convirtieron desde el formato matricial de la hoja de cálculo a formato EAV (*entity-attribute-value*), con una única observación almacenada en cada registro, puesto que este formato es más adecuado para el análisis automatizado de datos a pesar de ser menos intuitivo de visualizar (Wickham, 2007, 2014).

La base de datos resultante es la representación de un grafo orientado, donde los participantes son los nodos, y las distintas evaluaciones los arcos, de manera que los 16 participantes, evaluando 8 preguntas de 2 distintos participantes y a sí mismos, resultan en una tabla con 384 registros (16 x 8 x 3) que contiene los datos de cada arco (evaluador, evaluado, pregunta y valoración), pudiendo algunos de los datos ser valores nulos. Esta topología es complicada de gestionar en una base de datos relacional empleando SQL (*Structured Query Language*), de manera que el tratamiento de datos se hizo con el lenguaje estadístico R versión 3.2.0 (R Core Team, 2015), puesto que frente al paradigma declarativo de SQL permite la programación imperativa y la definición de funciones *ad-hoc*. Esta estrategia permitió manipular los datos para obtener agregados (máximos, mínimos y medias) tanto para las valoraciones recibidas como para las otorgadas utilizando la estrategia (Wickham, 2011) de segmentar los datos, aplicar una función de agregación a los fragmentos y finalmente recombinar los resultados (*split-apply-combine*) y también controlar su visualización (Wickham, 2009, 2010). La explotación de los resultados requirió aproximadamente 1000 líneas de código.

También posibilitó transformar la base de datos para invertir el grafo orientado de relaciones y obtener una tabla con las relaciones inversas, es decir, en lugar de la relación entre evaluador y las valoraciones recibidas, almacenar la relación entre éste y las valoraciones efectuadas.

#### A. Relación entre auto-evaluación y valoraciones recibidas y efectuadas

Uno de los objetivos del experimento era conocer la relación entre la auto-evaluación y las valoraciones recibidas de otros los participantes. Se esperaba observar una relación entre la percepción del estudiantado de su propio trabajo y la percepción de sus dos evaluadores elegidos aleatoriamente, tanto en las preguntas de evaluación libre como en las preguntas evaluadas mediante rúbrica, con la hipótesis de que la rúbrica propiciara resultados más coherentes al ser una evaluación más objetiva.

Paradójicamente, independientemente del uso o no de una rúbrica, la auto-evaluación resultó ser un mal predictor de la evaluación recibida (Fig. 1) no siendo la relación estadísticamente significativa en ambos casos (Tabla 1). Sin embargo, utilizando la auto-evaluación como predictor de las valoraciones realizadas si se apreció clara una relación (Fig. 2), siendo esta relación estadísticamente significativa tanto para la evaluación libre como para la evaluación mediante el uso de una rúbrica (Tabla 1).

Adicionalmente, representado la media de las valoraciones recibidas a través del tamaño del punto en el diagrama de dispersión que relaciona la auto-evaluación con las evaluaciones efectuadas (Fig. 2), se observó que, a igualdad de auto-evaluación, las valoraciones otorgadas que están por debajo de lo predicho por el modelo reciben sistemáticamente mejores valoraciones, cosa que hace pensar a los autores que un modelo de regresión múltiple podría explicar mejor los resultados en futuras investigaciones.

#### B. Incidencia del uso de rúbrica

Para el caso de la evaluación mediante el uso de una rúbrica de los trabajos, se esperaba una mayor consistencia en los resultados respecto a la valoración libre. Para explorar esta hipótesis, se representaron los resultados individuales utilizando un “*multiway dot plot*” (Cleveland, 1993), ubicando cada participante (ordenado según su auto-evaluación) en una fila, e indicando su auto-evaluación con un círculo y los rangos de las evaluaciones recibidas (en verde, encima) y de las evaluaciones realizadas (en naranja, debajo) como dos barras horizontales paralelas. De esta manera fue posible comparar ambos rangos para cada participante y al mismo tiempo relacionarlos con su auto-evaluación (Figs. 3 y 4).

De esta manera pudo apreciar la desconexión entre percepción propia y la percepción de los dos evaluadores de un mismo trabajo, así como la disparidad de las valoraciones de ambos tanto para el caso de la evaluación libre (Fig. 3) como en el caso de la evaluación mediante una rúbrica (Fig. 4), aunque en el segundo caso la dispersión fue menor, destacando dos apartados de la rúbrica (mapas y redactado) que mostraron resultados mucho más coherentes que los otros cuatro.

En contraste, se observó una mayor coherencia entre la auto-evaluación y las valoraciones efectuadas por cada participante, tanto en la evaluación libre (Fig. 3) como con el uso de una rúbrica (Fig. 4), teniendo la relación un mayor coeficiente de determinación (Tabla 2) para el caso de la valoración con

rúbrica (0,60) que en la valoración libre (0,34). Con el uso de la rúbrica se observa también una coherencia mayor en todos los apartados, siendo también los apartados de mapas y redactado los más coherentes.

#### C. Comparación de los resultados

A continuación se resumen los resultados de los modelos de regresión (Tabla 1), considerando la hipótesis nula  $H_0$  que la pendiente  $\beta$  es igual a cero, siendo  $p$  la probabilidad de que con los datos obtenidos,  $H_0$  sea cierta (la pendiente no tiene unidades puesto que es el cociente de dos magnitudes medidas en la misma unidad).

Tabla 1: Significancia estadística de la relación (pendiente de la línea de regresión) entre la auto-evaluación y a las valoraciones recibidas y efectuadas, en la valoración libre y en la valoración mediante rúbrica

Variables	Sin rúbrica	Con rúbrica
Auto-evaluación y valoraciones recibidas	$p = 0,99$ $> 0,05$	$p = 0,7730$ $> 0,05$
Auto-evaluación y valoraciones efectuadas	$p = 2,89e-05$ $< 0,05$	$p = 5,62e-05$ $< 0,05$

Como se puede apreciar, no es significativa la relación entre la auto-evaluación y las valoraciones obtenidas, mientras que sí lo es la relación entre la auto-evaluación y las valoraciones efectuadas, en ambos casos independientemente del uso de la rúbrica.

En cuanto a los coeficientes de determinación  $R^2$  (Tabla 2) se puede apreciar que para la relación entre auto-evaluación y valoraciones recibidas, los coeficientes son cercanos a cero tanto en la evaluación libre como mediante el uso de la rúbrica. Para el caso de la relación entre auto-evaluación y valoraciones efectuadas, el uso de la rúbrica incrementa apreciablemente este valor.

Tabla 2: Coeficiente de determinación  $R^2$  entre la auto-evaluación y a las valoraciones recibidas y efectuadas, en la valoración libre y en la valoración mediante rúbrica

Variables	Sin rúbrica	Con rúbrica
Auto-evaluación y valoraciones recibidas	$R^2 = 4,2e-06$	$R^2 = 4,7e-03$
Auto-evaluación y valoraciones efectuadas	$R^2 = 0,34$	$R^2 = 0,60$

Cabe enfatizar que estos datos no corresponden a dos experimentos distintos sino que corresponden a información referida a un mismo hecho (la valoración de un trabajo ajeno), es decir, cada vez que un participante hizo una valoración, se generaron dos datos: (a) una de las dos valoraciones que recibió el trabajo ajeno evaluado, y (b) una de las dos valoraciones realizadas por el participante. Estos dos datos fueron los que se compararon con la auto-evaluación de cada participante.

También cabe destacar que la auto-evaluación y la valoración de trabajos ajenos se hicieron en el orden y en el momento que cada estudiante estimó oportuno, y por lo tanto este hecho quedó fuera del control del experimento.

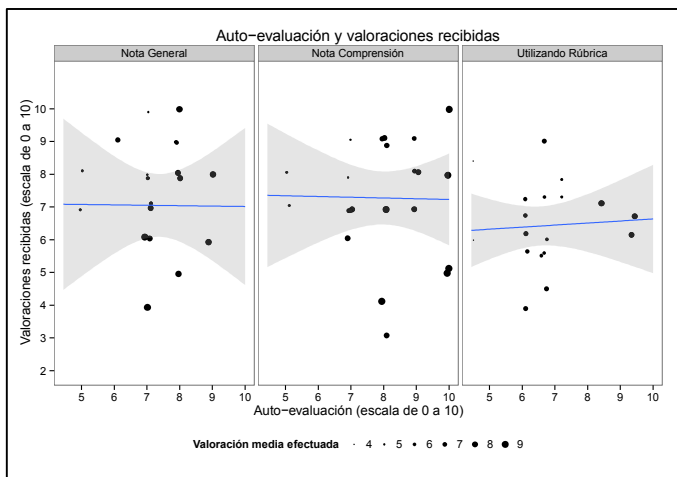


Figura 1: Diagrama de dispersión entre auto-evaluación y valoraciones recibidas. Respuestas libres (izquierda y centro) y utilizando rúbrica (derecha). Posición de los puntos con "jitter" y dimensión proporcional a la valoración otorgada.

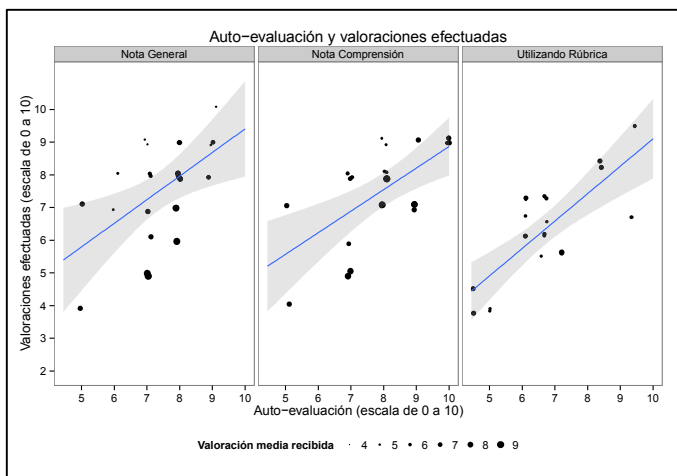


Figura 2: Diagrama de dispersión entre auto-evaluación y valoraciones otorgadas. Respuestas libres (izquierda y centro) y utilizando rúbrica (derecha). Posición de los puntos con "jitter" y dimensión proporcional a la valoración recibida.

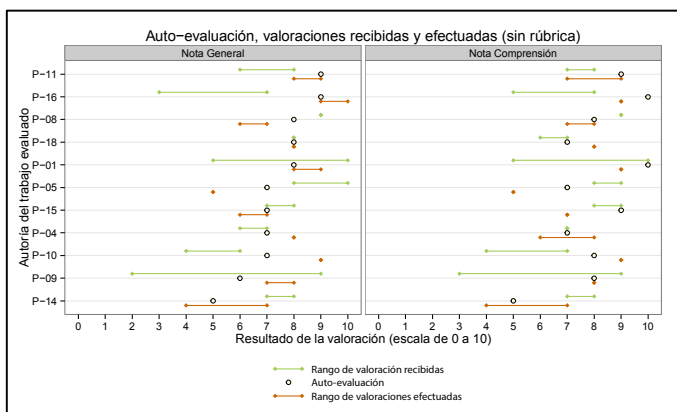


Figura 3: "Multiway dot plot" de la auto-evaluación (círculo) y rango de valoraciones recibidas (verde) y otorgadas (naranja) de las dos preguntas sin el uso de rúbrica. Orden según valoración propia.

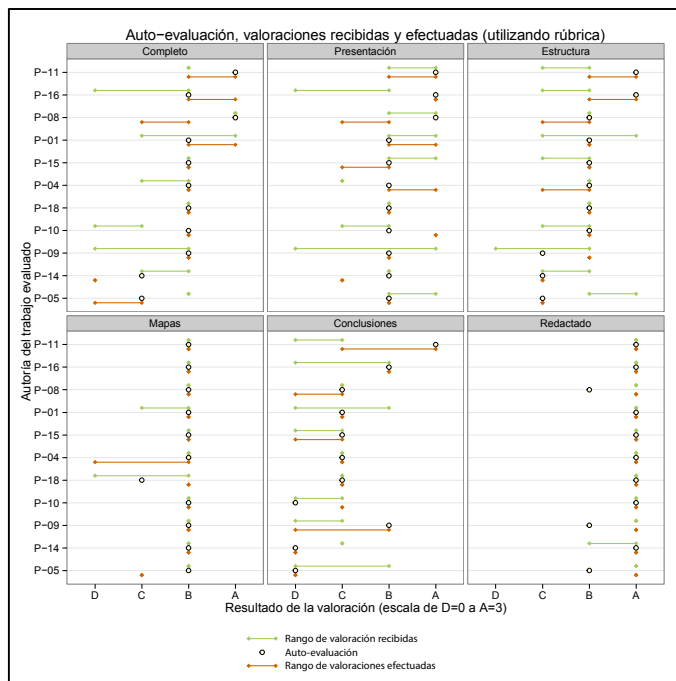


Figura 4: "Multiway dot plot" de la auto-evaluación (círculo) y rango de valoraciones recibidas (verde) y otorgadas (naranja) de las seis preguntas mediante rúbrica. Orden según valoración propia.

#### D. Concordancia de los distintos apartados de las evaluaciones mediante rúbrica

Se calcularon las correspondientes tablas de contingencia (4x4 evaluaciones posibles) para cada una de las 6 apartados de la rúbrica, tanto para la relación de la auto-evaluación con la evaluación recibida como con la realizada. Para representar las relaciones entre las distintas evaluaciones, se descartó utilizar un diagrama de cribado ("sieve diagram"), de uso habitual para representar relaciones entre variables categóricas (Friendly & SAS Institute, 2001; Meyer, Zeileis, & Hornik, 2006), debido a que las categorías (evaluaciones) eran de tipo ordinal (insatisfactorio, cumple mínimos, satisfactorio, o excepcional). Se consideró más adecuado utilizar un diagrama de acuerdo ("agreement chart"), donde se visualizan los valores observados en la diagonal de una matriz de confusión como rectángulos oscuros, respecto a los valores esperados en colores claros (Bangdiwala & Shankar, 2013).

En el diagrama (Fig. 5) se representaron dos casuísticas para cada uno de los apartados de la rúbrica: (a) el acuerdo entre la auto-evaluación y la evaluación recibida, y (b) el acuerdo entre la auto-evaluación y la evaluación realizada. Los seis apartados analizados fueron:

- Compleción (A1)
- Presentación (A2)
- Estructura (A3)
- Salidas gráficas (A4)
- Conclusiones (A5)
- Redactado (A6)

Con el diagrama se puede apreciar que el estudiantado tiende a valorar en la franja alta de las calificaciones, y que las valoraciones propias son más consistentes con las realizadas

que con las recibidas. Además de realizar la visualización se calculó la correspondiente estadística Bangdiwala B para la relación entre la auto-evaluación y las valoraciones recibidas y realizadas (Tabla 3) que tiene un rango entre 0 (desacuerdo) y 1 (máximo acuerdo). El máximo acuerdo se observó en los apartados referentes a mapas (A4) y redactado (A6).

Tabla 3: Resultados del cálculo de Bangdiwala B

Relación	A1	A2	A3	A4	A5	A6
Recibidas	0,28	0,32	0,33	0,90	0,31	0,68
Realizadas	0,62	0,41	0,51	0,79	0,58	0,70

## 5. CONCLUSIONES

A la luz de los resultados, los autores creen que la explicación de la disparidad entre las evaluaciones recibidas y la auto-evaluación, así como la similitud entre esta y las evaluaciones realizadas se deben a un mecanismo similar al efecto Dunning-Kruger. La conjetura de los autores es que al ser los participantes legos en la materia, como se comprobó en las encuestas de perfil de estudiantado, se encuentran en una fase en la que les es complicado evaluar correctamente el trabajo de otros, aún con el uso de una rúbrica, y por lo tanto reflejan la percepción de su propio conocimiento en los trabajos que evalúan, considerando que su grado de conocimiento es similar al de sus compañeros y compañeras.

Para comprobar esta hipótesis, los autores planean hacer el mismo ensayo con estudiantes de máster, y comprobar si los mayores conocimientos y experiencia de estos mitigan este efecto, acercando las valoraciones recibidas a las propias y alejando de estas las valoraciones realizadas. También queda fuera del ámbito de este documento la comparación de las evaluaciones del profesorado con las realizadas por el alumnado y la evaluación de la experiencia de usuario del estudiantado, especialmente en los aspectos relacionados con la motivación.

Por las razones expuestas en este artículo, la respuesta a la pregunta de investigación acerca de la utilidad de la evaluación entre iguales es que, a la luz de los resultados obtenidos, la evaluación entre iguales no es adecuada para la evaluación del alumnado, en particular para el caso de estudiantado novel, independientemente del uso o no de una rúbrica, hecho que no invalida los valores que puede aportar, como herramienta educativa, la revisión crítica de trabajos propios o ajenos.

Finalmente, también se pone de manifiesto la necesidad de elaborar una rúbrica más detallada que elimine en la medida que sea posible la ambigüedad en la evaluación de los trabajos para futuras ediciones del curso.

## AGRADECIMIENTOS

La presente investigación se ha hecho en el contexto del proyecto EDU2012-37247, E-LEARNING 3.0 EN LA DOCENCIA DE LA ARQUITECTURA. CASOS DE ESTUDIO DE INVESTIGACIÓN EDUCATIVA PARA UN FUTURO INMEDIATO.

Los autores quieren también agradecer los comentarios de los evaluadores anónimos.

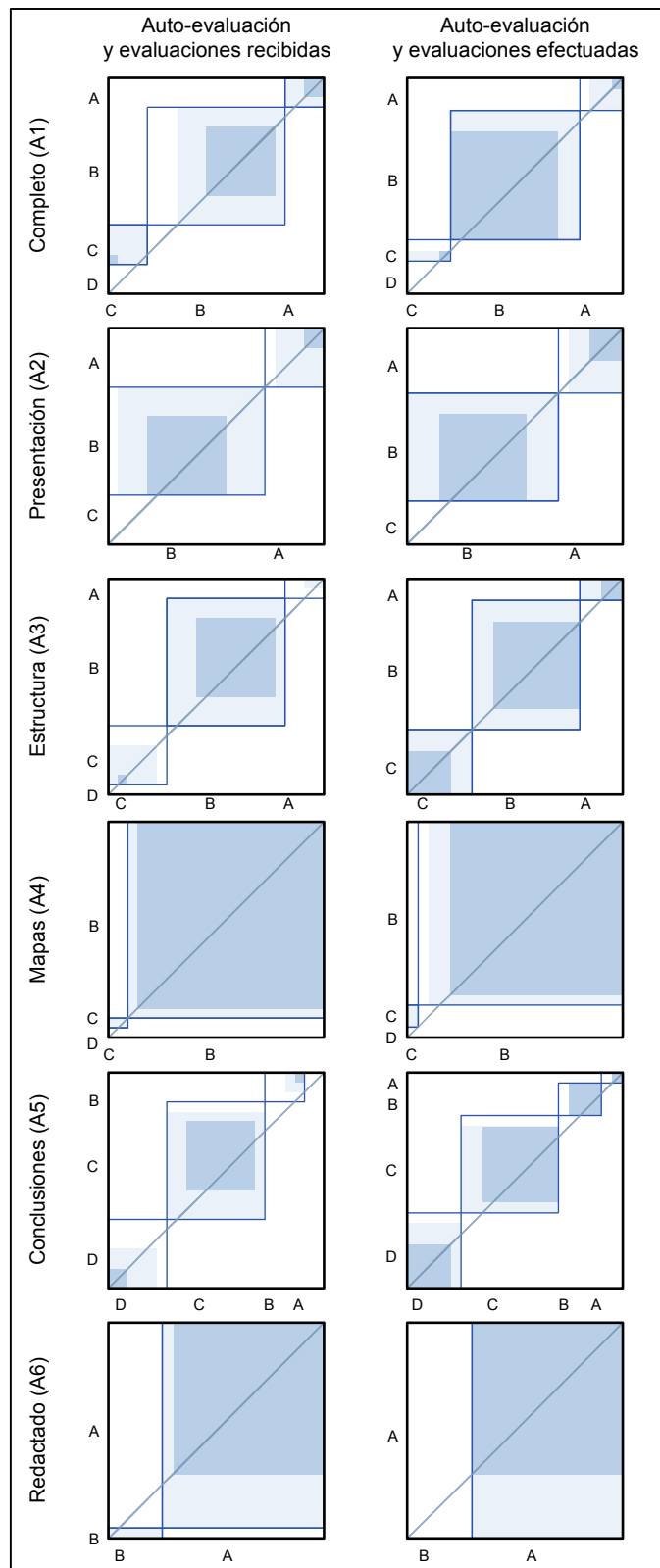


Figura 5: Diagrama de acuerdo de la auto-evaluación (el eje horizontal) y las evaluaciones recibidas (izquierda) y realizadas (derecha) en el eje vertical para cada uno de los apartados de la rúbrica (A1-A6). Acuerdo exacto en azul oscuro y acuerdo parcial en azul claro.

REFERENCIAS

- Ames, D. R., & Kammrath, L. K. (2004). Mind-Reading and Metacognition: Narcissism, not Actual Competence, Predicts Self-Estimated Ability. *Journal of Nonverbal Behavior*, 28(3), 187–209. <http://doi.org/10.1023/B:JONB.0000039649.20015.0e>
- Bangdiwala, S. I., & Shankar, V. (2013). The agreement chart. *BMC Medical Research Methodology*, 13(1), 97. <http://doi.org/10.1186/1471-2288-13-97>
- Cleveland, W. S. (1993). *Visualizing Data* (1 edition). Murray Hill, N.J.: Summit, N.J.: Hobart Press.
- Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, 84(1), 5–17. <http://doi.org/10.1037/0022-3514.84.1.5>
- Friendly, M., SAS Institute. (2001). *Visualizing categorical data*. Cary, NC: SAS Institute.
- García-Almirall, P., Redondo, E., Valls, F., & Corso, J. M. (2014). Experiencia docente en la enseñanza de Sistemas de Información Geográfica en Arquitectura. In *9th Iberian Conference on Information Systems and Technologies* (Vol. 1, pp. 407–412). Barcelona, Spain. <http://doi.org/http://dx.doi.org/10.1109/CISTI.2014.6876870>
- García-Almirall, P., Redondo, E., Valls, F., & Fonseca, D. (2015). Evaluación entre iguales en la enseñanza de Arquitectura: Aplicación en la enseñanza de Sistemas de Información Geográfica. In Á. Rocha, A. Martins, G. Paiva Dias, L. P. Reis, & M. Pérez Cota (Eds.), *10th Iberian Conference on Information Systems and Technologies* (Vol. I, pp. 1009–1015). Aveiro, Portugal.
- Google. (2015). Google Forms. Retrieved February 10, 2015, from <http://www.google.com/forms/>
- Kruger, J., & Dunning, D. (1999). Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology*, 77, 1121–1134. <http://doi.org/10.1037/0022-3514.77.6.1121>
- Meyer, D., Zeileis, A., & Hornik, K. (2006). The Strucplot Framework: Visualizing Multi-way Contingency Tables with vcd. <http://www.jstatsoft.org/v17/i03/paper>.
- R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from <http://www.R-project.org/>
- Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, 21(12), 1–20. Retrieved from <http://www.jstatsoft.org/v21/i12>
- Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Dordrecht, New York: Springer.
- Wickham, H. (2010). A Layered Grammar of Graphics. *Journal of Computational and Graphical Statistics*, 19(1), 3–28. <http://doi.org/10.1198/jcgs.2009.07098>
- Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, 40(1), 1–29. Retrieved from <http://www.jstatsoft.org/v40/i01>
- Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, 59(10), ??–?? Retrieved from <http://www.jstatsoft.org/v59/i10>