UNIVERSITAT POLITÈCNICA DE VALÈNCIA | IN-RED 2021 VII Congreso de Innovación Educativa y Docencia en Red

Congreso In-Red 2021
UPV, 13, 14 y 15 de julio de 2021
Doi: http://dx.doi.org/10.4995/INRED2021.2021.13810

Exámenes en grupo y pruebas de corrección como alternativas a la evaluación

Miguel Rebollo[1]

[1]Universitat Politècnica de València

Abstract

This work shows evaluation strategies for objective tests that incorporate elements that allow students to reflect on their learning during the trial. We have introduced two different methods in one of the partial exams of the subject in which we have applied the innovation. The first consisted of using peer correction as an evaluation element instead of the students' answers to a series of multiple-choice questions. The second one consisted of group resolution of an exercise. To evaluate the results, we compared the results obtained in the first partial exam, the results obtained in equivalent tests in previous years, and the answers from participants to a validated questionnaire on new exam modalities.

Keywords*: evaluation, higher education, objective tests, questionnaires, self-perception, academic achievement*

Resumen

Este trabajo muestra el uso de estrategias de evaluación para pruebas objetivas que incorporan elementos que permitan al alumnado reflexionar sobre su propio aprendizaje durante la realización de la prueba. Se han planteado dos métodos distintos en uno de los parciales de la asignatura en la que se ha aplicado la innovación. El primero ha consistido en emplear la corrección por pares como elemento de evaluación en lugar de las respuestas propias a una serie de preguntas de respuesta múltiple. La segunda ha consistido en la resolución en grupo de un ejercicio. Para valorar los resultados se ha comparado con los resultados obtenidos en el primer parcial, los resultados obtenidos en pruebas equivalentes en cursos anteriores y las respuestas de los participantes en un cuestionario validado sobre nuevas modalidades de exámenes.

Keywords*: evaluación, pruebas objetivas, cuestionarios, autopercepción, rendimiento académico*

1 Introducción

La evaluación es una de las mayores preocupaciones de docentes y estudiantes. Los primeros intentando garantizar que los resultados obtenidos corresponden con las competencias adquiridas a lo largo de la asignatura y los segundos para conseguir un expediente que luego les permita incorporarse con facilidad al mercado laboral.

Existen una gran variedad de formas de evaluar al alumnado y habitualmente se movilizan varias de ellas durante el desarrollo de la docencia. El peso de pruebas objetivas escritas es cada vez menor y se reemplazan por otro tipo de demostraciones más cercanas a la evaluación auténtica (J. Herrington 2006; Maina 2004; Jan Herrington y Oliver 2000) en la que se deben movilizar todas las competencias para resolver situaciones del mundo real.

Sin embargo, las pruebas objetivas siguen siendo útiles y resultan ser un elemento para testear de forma rápida la adquisición de unos conocimientos teóricos mínimos o incluso como elemento de autoevaluación para comprobar la progresión y el grado de comprensión de los conceptos Fuentes y Beltran-Sanchez 2020.

Existen formas de introducir elementos de reflexión en los cuestionarios para que la prueba incorpore un carácter formativo. De esta manera, también se consigue que se trabaje la autopercepción del propio aprendizaje (Chevalier y col. 2009). Una de ellas es la inclusión de la justificación de las repuestas en pruebas de respuesta múltiple (Germain y col. 2016), que es una de las aproximaciones que emplearemos en el presente trabajo.

La realización de exámenes en dos etapas se ha estudiado ampliamente, si bien la segunda de ellas tiende a ser una actividad grupal Zipp 2007; Levy, Svoronos y Klinger 2018. En esos casos, el examen deja de ser un elemento de evaluación exclusivamente sumativa y se convierte en un elemento formativo. El alumnado reconoce que este tipo de pruebas dan un mayor soporte al aprendizaje y restan estrés a las pruebas tradicionales.

Uno de los problemas en la realización de las pruebas individuales es la copia entre pares. Esta problemática se ha acrecentado en la situación del curso 2020-2021, en la que una parte importante de la docencia se ha desplazado a entornos en línea junto con la evaluación. Una de las soluciones habituales es la vigilancia para impedir las copias. Sin embargo, existe otra estrategia que consiste en, en lugar de impedir la colaboración en la realización de las pruebas, fomentarlas y usarlas de forma que el trabajo individual no se diluya por completo y siga teniendo un peso importante. Una de ellas es la realización de exámenes en grupo (Molina Jordá 2016; Bloom 2009; Mahoney y Harris-Reeves 2019). Se ha demostrado que el rendimiento obtenido es superior y los resultados sugieren que promueven aprendizajes de orden superior, independientemente del desempeño de los estudiantes.

El meta-análisis de Zell y colaboradores realizado sobre el «efecto mejor que la media» (BTAE—*better-than-average-effect*—) muestra que, en general, los estudiantes tienen a creer que son mejores que la media. Esto podría hacer que, en este tipo de pruebas, el alumnado se muestre reacio por considerar que trabajar en grupo le va a perjudicar E y col. 2020.

Sin embargo, los resultados de Moore, que ha estudiado la percepción del alumnado acerca de los exámenes colaborativos, no muestran ese efecto Moore 2010. La mayoría nunca había tenido una experiencia previa con esa forma de evaluar. Identificaron cuatro ventajas: la discusión aumenta la comprensión, tienen una oportunidad de obtener mejor calificación, es una oportunidad de trabajar en equipo y mejora la responsabilidad individual. A cambio, hay estudiantes que son reacios a

depender del esfuerzo de otros y pueden surgir conflictos por la presión de estar realizando un examen.

2 Objetivos

Los objetivos que se pretende conseguir introduciendo una serie de cambios en las pruebas objetivas son los siguientes:

- incluir en las pruebas una parte de evaluación formativa que permita la reflexión sobre autoaprendizaje
- explorar alternativas a las pruebas objetivas de respuesta múltiple
- incorporar mecanismos que dificultan la copia en pruebas en línea (secundario)

3 Desarrollo de la innovación

3.1 Participantes

La propuesta se plantea en la asignatura «Informática Aplicada» (IA), en el Grado en Gestión y Administración de Empresas (GAP) que se imparte en la Facultad de Administración y Dirección de Empresas (FADE) de la Universitat Politècnica de València (UPV). Se trata de una asignatura obligatoria de primer curso que se imparte en el primer semestre.

En el curso 2020-21 están matriculados 88 estudiantes, todos los cuales han participado en la innovación ($n = 88$). Para comparar los resultados obtenidos, se ha extraído información de cursos anteriores, recogiendo un total de datos de 329 estudiantes.

Tabla 1: Número total de participantes en el estudio

curso	n
2017-18	86
2018-19	80
2019-20	75
2020-21	88
total	329

3.2 Formato de las sesiones

La asignatura tiene 4,5 créditos que se reparten durante 15 semanas, estructurados en 1,5 horas teóricas y 2 horas de seminarios o trabajo de laboratorio. Los resultados del presente trabajo corresponden a la evaluación de los contenidos teóricos de la asignatura. Hay dos actos de evaluación: uno a mitad del semestre y un segundo al finalizar.

La evaluación de la asignatura combina distintos sistemas de evaluación con el fin de reunir información suficiente para poder tomar una decisión informada sobre la calificación final, incorporando pruebas de disto carácter (formativas y sumativas), en distintos momentos (inicial, continua y final), de distinta procedencia (individual y grupal), de distintas fuentes (profesorado, autoevaluación y por pares) y empleando distintas técnicas.

A modo de resumen, se puede considerar con la calificación final del curso se obtiene a partir de tres fuentes: (i) el trabajo realizado en el aula, incluyendo en este apartado también las pruebas propuestas al finalizar cada parte de los contenidos y las actividades a realizar en el tiempo de estudio personal; (ii) las pruebas objetivas escritas para los contenidos teóricos y (iii) un portafolios de trabajo para la parte práctica de la asignatura. El trabajo en el aula tiene un peso del 30 % en la calificación final y está formado principalmente por muestras grupales. Las pruebas objetivas suponen otro 30 % de la calificación (dividido en dos partes) y el portafolios supone el 40 % restante y, aunque su elaboración es personal, parte de su peso corresponde a evaluación por pares, además de su evaluación formativa a lo largo de su elaboración.

Como se ha mencionado, en el presente trabajo se analiza el efecto que tienen los cambios en el planteamiento de las pruebas objetivas. En concreto, en una de las dos pruebas parciales, con un peso del 15 % en la calificación final. Aunque la prueba tiene relativamente bajo impacto en la calificación final, por tratarse de una prueba parcial en condiciones de examen, el alumnado tiene una percepción de una relevancia considerablemente más alta de la real.

3.3 Estructura de las pruebas objetivas

Las pruebas objetivas se realizan sobre los contenidos teóricos de la asignatura. Una primera prueba incluye los tres primeros temas (introducción a las computadoras y conceptos básicos de hardware) y se realiza en el mes de noviembre, aproximadamente a la mitad del semestre. Está formada por una serie de preguntas de respuesta múltiple (cuatro opciones, solo una correcta) y una parte de preguntas cortas de desarrollo. La segunda prueba se realiza al finalizar el semestre sobre los cuatro temas restantes (conceptos básicos de software, redes e internet). Tiene la misma estructura, sustituyendo las preguntas por la resolución de un ejercicio de diseño de bases de datos. La innovación se ha realizado en esta segunda prueba, usando la primera prueba como control sobre el mismo grupo de estudiantes.

La prueba objetiva se dividió en dos partes separadas. La primera afecta a las preguntas de respuesta múltiple. Para evaluar al alumnado se plantea como una prueba de corrección: a cada estudiante se le proporciona un examen resuelto y su calificación se obtiene de la valoración razonada de las correcciones de las respuestas que se le han facilitado.

Dada la dificultad de realizarlo con las propias pruebas, se plantea de la siguiente forma. Se construyen dos modelos de examen con 21 preguntas, siete de cada tema, con cuatro opciones de las cuales solo una es correcta. Sobre cada modelo, se generan ocho enunciados desordenando tanto preguntas como opciones (en total 16 enunciados distintos). Para simular las respuestas al examen,

se generar tantas soluciones al azar como estudiantes ($n = 88$). A cada estudiante se le asigna uno de los 16 enunciados y una solución. La respuesta al examen consiste en, para cada pregunta

1. marcar si la solución proporcionada es correcta o no
2. justificar la decisión

La justificación es obligatoria y no puede ser la negación del enunciado o señalar la respuesta correcta: se debe indicar un motivo por el que la opción no es correcta. Al ser respuestas al azar, es más probable que estas sean falsas. Pero puesto que la puntuación corresponde a la justificación y no a la respuesta en sí, consideramos que no es un factor que haya afectado a las calificaciones finales.

> *Ejemplo: la pregunta 1 del enunciado es*
>
> 1. *la memoria que emplea el ordenador para su arranque es la memoria*
>
> a) *RAM*
>
> b) *ROM*
>
> c) *cache*
>
> d) *USB*
>
> *y en la hoja de respuestas aparece marcada* 1,a
>
> *En la pregunta 1 se debe marcar si la respuesta correcta a esa pregunta es la a) En este caso no lo es, así que hay que marcar* «incorrecta» *y justificarlo con algo como* «La RAM es la memoria que usa el ordenador para la ejecución de los programas»

La segunda parte de la prueba consiste en un ejercicio de diseño de bases de datos a partir de un enunciado. En este caso, se optó por la resolución en grupo. Se formaron grupos al azar de entre tres y cuatro personas. Cada grupo trabajó sobre la solución y luego cada persona elaboró su propio diseño para la entrega. Se permitía entregar directamente la solución del grupo si pensaban que era la correcta, o presentar una versión modificada si se consideraba que la solución consensuada por el grupo no era la correcta.

Inicialmente, para evitar responsabilidades asimétricas en el grupo y que una o dos personas se encargaran de la solución o se dividieran la tarea, se plantea que la nota para el grupo será la de uno de sus miembros escogido al azar. De esta forma, es responsabilidad del grupo que todos sus integrantes hayan comprendido la solución y sean capaces de replicarla en sus respuestas. Sin embargo, puesto que casi todo el alumnado entregó versiones propias de los ejercicios, se optó por emplear la nota personal directamente al considerar que la gran mayoría había hecho alguna aportación propia o modificación a la respuesta del grupo.

4 Resultados

4.1 Reacción de los participantes

Para validar los resultados, se ha utilizado una encuesta validada sobre la aceptación de nuevos mecanismos de evaluación (Leeming 2002). El cuestionario está compuesto por 6 preguntas a las que se podía responder con si/no/no lo sé. Se planteó el mismo cuestionario para las dos innovaciones por separado. El cuestionario se pasó una vez publicadas las notas. Se obtuvieron $m = 15$ respuestas. Con esta muestra sobre la población total ($n = 88$), el error muestral obtenido para un intervalo de confianza del 95 % es del 22 %.

La Tabla 2 recoge los resultados obtenidos en el cuestionario. Respecto a la prueba planteada como un ejercicio de corrección, el procedimiento tiene una percepción claramente negativa en todos los aspectos valorados. El sentimiento mayoritario es de escepticismo inicial al plantear el mecanismo de la prueba y la no recomendación de este sistema para cursos siguientes. Aunque en general la se prefiere los test tradicionales frente a este método, hay una ligera diferencia en las dos preguntas que lo plantean en el sentido de que si pudieran escoger no se decantarían por el tipo de prueba planteado.

En cuanto al ejercicio en grupo, la percepción es la contraria. En general hay una mayor aceptación a este tipo de pruebas y las opiniones no están tan decantadas hacia uno u otro extremo. No hay una percepción clara de que hayan aprendido más con este tipo de examen. Probablemente esta respuesta esté indicando también la consideración de los exámenes como pruebas de diagnóstico y no como oportunidades de aprendizaje.

Tabla 2: Respuestas al cuestionario de nuevos métodos de evaluación

Ejercicio de corrección	Sí	No	No lo sé
Era escéptica/o cuando se propuso el método	67 %	7 %	27 %
He aprendido más que si hubiera tenido un test convencional	33 %	40 %	27 %
Si me dan elegir, prefiero este sistema	20 %	67 %	13 %
Prefiero un examen de corrección a un test convencional	40 %	53 %	7 %
El examen de corrección ha sido una experiencia horrible	40 %	47 %	13 %
Recomiendo el examen de corrección para el próximo semestre	7 %	60 %	33 %
Ejercicio en grupo	**Sí**	**No**	**No lo sé**
Era escéptico/a cuando se propuso el método	33 %	60 %	7 %
He aprendido más que si hubiera tenido un examen individual	47 %	40 %	13 %
Si me dan elegir, prefiero este sistema	60 %	20 %	20 %
Prefiero un examen en grupo a uno individual	60 %	20 %	20 %
El examen en grupo ha sido una experiencia horrible	13 %	73 %	13 %
Recomiendo el examen en grupo para el próximo semestre	53 %	33 %	13 %

4.2 Puntuaciones obtenidas

Por otro lado, se han estudiado las calificaciones obtenidas en los actos de evaluación con l fin de observar el rendimiento académico de los estudiantes. Se compara en primer lugar con los resultados obtenidos por los mismos individuos en el primer parcial (poblaciones idénticas), que fue un examen tradicional. Por otro lado, para valorarlo frente a los mismos contenidos, también se compara con las calificaciones obtenidas sobre los mismos contenidos en cursos anteriores (poblaciones diferentes).

Tabla 3: Comparación de los resultados del primer y segundo parcial

curso	parcial 1	parcial 2	p-valor ($< 0{,}05$)	t. efecto ($> 0{,}8$)
2017-18	$4,6 \pm 1,6$	$5,4 \pm 1,6$	0,0002	0,65
2018-19	$4,2 \pm 1,5$	$3,5 \pm 1,8$	1e-5	0,61
2019-20	$3,5 \pm 1,6$	$5,0 \pm 1,7$	8,3e-11	1,24
2020-21	$4,9 \pm 1,7$	$6,8 \pm 1,9$	1,7e-6	1,5

En la comparación con el primer parcial se considera la calificación global de cada prueba, ya que la estructura de las mismas, el tipo de preguntas y la extensión es distinta. Los resultados muestran una clara mejoría del segundo parcial ($M = 6,8$ $SD = 1,9$) respecto al primero ($M = 4,9$ $SD = 1,7$), siendo $t(81) = 4,9$ con p-valor $p = 1,8e-6$ y tamaño del efecto $ES = 1,5$. Para comprobar la relación entre las calificaciones del primer y segundo parcial en cursos anteriores se han realizado los cálculos con los datos disponibles. Los resultados obtenidos aparecen en la Tabla 3.

Para valorar el rendimiento académico sobre los mismos contenidos, se ha realizado un ANOVA sobre los valores obtenidos en las preguntas de respuesta múltiple y los ejercicios de diseño de bases de datos. Puesto que el número de preguntas y la valoración de cada parte presenta ligeras diferencias en distintos cursos, se ha optado por normalizar las calificaciones en el intervalo [0,1].

Para el planteamiento de la prueba como un examen de corrección, se obtiene un valor $F = 14,54$ comparando los valores en cada curso, resultado que tiene una probabilidad de $p = 6,4e-9$, lo que nos lleva a rechazar la hipótesis de igualdad de medias. De la misma forma, la resolución del ejercicio de bases de datos en grupo se ha comparado con la resolución individual en cursos anteriores. En este caso, se obtiene un valor $F = 6,56$, que sigue teniendo una probabilidad $p = 0,0002$ que lleva a rechazar la hipótesis de igualdad de medias. Sin embargo, en la Figura 1 podemos apreciar que la diferencia se encuentra en el curso 2018 en ambos casos. Este hecho lo podemos corroborar con el tamaño del efecto, que es $ES = 1{,}4$ para el examen de corrección y de $ES = 0,72$ (cercano al límite de 0,8) para la resolución grupal. En los demás casos, no hay evidencia de una diferencia significativa entre los resultados obtenidos.

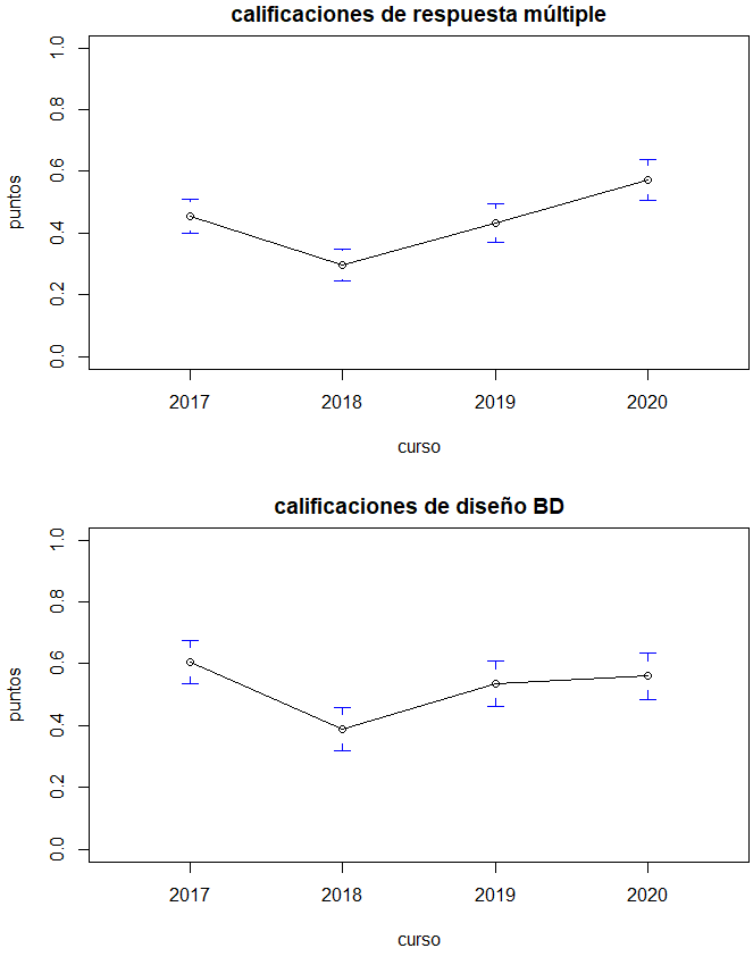

Fig. 1: ANOVA: Intervalos de confianza de las calificaciones del segundo parcial en cada curso

5 Discusión y conclusiones

La percepción del alumnado acerca de las dos innovaciones introducidas en la evaluación es clara: hay una sensación de rechazo ante plantear la prueba de respuesta múltiple como una prueba de corrección y de aceptación a las pruebas grupales. La realización del cuestionario es posterior a la publicación de las calificaciones. Y si comparamos la percepción propia con los resultados obtenidos no se corresponde. Posiblemente se deba a que no se dispone de elementos comparativos ni con el resto de la clase ni con los resultados de cursos anteriores.

Los resultados muestran que la calificación del segundo parcial es claramente superior a la del primero (ver Figura 2), como pone de manifiesto el p-valor obtenido al analizar la diferencia de las medias. En las dos gráficas de la Figura 1 también se puede apreciar que la diferencia en el

test es la responsable del aumento de la calificación, lo que contradice la percepción de los y las estudiantes.

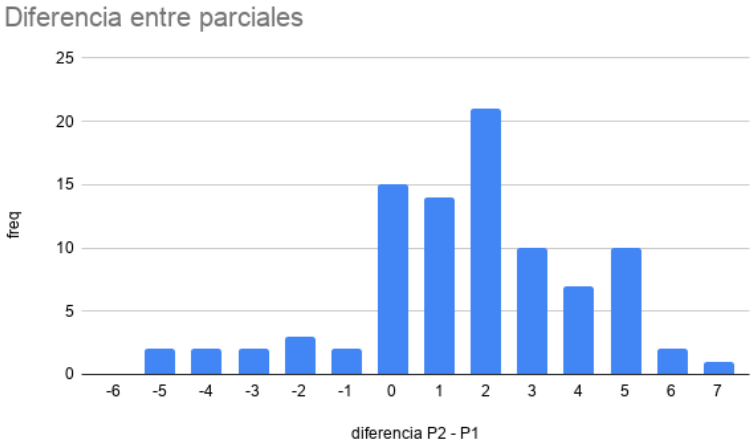

Fig. 2: Diferencia entre la calificación de ambos parciales (curso 2020)

El análisis cuantitativo muestra que no existe una diferencia significativa entre la calificación obtenida en el primer y segundo parcial comparada con la situación de cursos anteriores. Lo habitual es que la calificación del segundo parcial sea superior a la del primero (excepto en el curso 2018), y esa situación se repite. A pesar de que el p-valor nos indicaría un efecto significativo para todos los cursos, el tamaño del efecto nos sugiere que en los dos primeros (2017 y 2018) no hay realmente una diferencia clara entre el primer y el segundo parcial, que ya sí aparece en los dos últimos cursos. El que esta situación se repita nos indicaría que el cambio en la tipología del examen no ha introducido artefactos en las calificaciones del alumnado.

La comparación de las calificaciones del segundo parcial nos sugiere lo mismo: que los cambios planteados en el examen permiten evaluar correctamente los contenidos correspondientes sin que, en principio, se detecten desviaciones de los resultados esperados.

Algo a resaltar es que este curso sí que ha habido una situación especial: las pruebas han sido en línea por la suspensión de la actividad presencial por la pandemia del covid-19. Durante el curso se ha puesto de manifiesto la preocupación del profesorado por la validez de las pruebas realizadas a distancia. Los resultados obtenidos sugieren que esta alternativa es válida para evaluar al alumnado por medios en línea, sin que se hayan detectado copias masivas.

El planteamiento de la prueba de corrección ha permitido elaborar un examen individualizado para cada persona. El tiempo asignado a la prueba era de 2 minutos por cada pregunta y en ese tiempo había que determinar si la respuesta era correcta y dar una justificación. No había tiempo material para que se intercambiaran respuestas, además el porcentaje de preguntas comunes era muy bajo.

La realización de exámenes en grupo consigue que el alumnado se encuentre en la misma situación de cara al examen, sin que aquellas personas que tuvieran mejores contactos obtuvieran alguna ventaja. Como se ha comentado, la gran mayoría planteó su propia versión del ejercicio con ligeras

modificaciones sobre la resolución del grupo, lo que lleva a pensar que la forma de trabajar fue la correcta. La dispersión de las notas obtenidas en esta parte lo corrobora.

El principal inconveniente es el tiempo que requiere la corrección de este tipo de pruebas. No existen herramientas que permitan automatizar la corrección de los test. Y puesto que cada examen era único y la puntuación estaba basada en la justificación, el tiempo necesario para su revisión fue excesivo. Para su implantación en un entorno presencial, se sugiere

- utilizar las respuestas reales del alumnado en lugar de generarlas automáticamente
- reducir el número de exámenes distintos

Algo que se debería evaluar utilizando un estudio longitudinal a medio o largo plazo es si incluir la corrección, que obliga a una reflexión integrada en la propia prueba, consigue unos aprendizajes significativos, pero no existe ninguna otra asignatura en la que se extiendan los conceptos que aquí se desarrollan, por lo que habría que buscar algún otro mecanismo. También sería interesante conseguir que el alumnado perciba la mejora del rendimiento propio, ya que es un punto que ha pasado desapercibido.

6 Referencias

Referencias bibliográficas

Bloom, Davida (2009). "Collaborative Test Taking: Benefits for Learning and Retention". En: *College Teaching* 57.4, págs. 216-220.

Chevalier, Arnaud y col. (2009). "Students' academic self-perception". En: *Economics of Education Review* 28.6, págs. 716-727.

E, Zell y col. (2020). "The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis". En: *Psychol Bull* 146.2, págs. 118-149.

Fuentes, A. Hambleton y J.A. Beltran-Sanchez (2020). "Standardized objective exam vs. high fidelity simulation: two models, same content". En: *Proceedings of the 9th International Workshop on Innovative Simulation for Healthcare (IWISH 2020)*, págs. 13-19.

Germain, Francisco y col. (feb. de 2016). "Formulación de preguntas de respuesta múltiple: un modelo de aprendizaje basado en competencias". En: *FEM: Revista de la Fundación Educación Médica* 19, págs. 27-38.

Herrington, J. (2006). "Authentic E-Learning in higher education: Design principles for authentic learning environments and tasks". En: *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (ELEARN) 2006*, págs. 3164-3173.

Herrington, Jan y Ron Oliver (2000). "An Instructional Design Framework for Authentic Learning Environments". En: *Educational Technology Research and Development* 48.3, págs. 23-48.

Leeming, Frank C. (2002). "The Exam-A-Day Procedure Improves Performance in Psychology Classes". En: *Teaching of Psychology* 29.3, págs. 210-212.

Levy, Dan, Theodore Svoronos y Mae Klinger (2018). "Two-stage examinations: Can examinations be more formative experiences?" En: *Active Learning in Higher Education*, págs. 1-16.

Mahoney, John W y Brooke Harris-Reeves (2019). "The effects of collaborative testing on higher order thinking: Do the bright get brighter?" En: *Active Learning in Higher Education* 20.1, págs. 25-37.

Maina, F. (2004). "Authentic learning : perspectives from contemporary educators". En:

Molina Jordá, José Miguel (2016). *Test grupales como potenciadores del aprendizaje significativo.*

Moore, Lori L. (2010). "Students' Attitudes and Perceptions about the Use of Cooperative Exams in an Introductory Leadership Class". En: *Journal of Leadership Education* 9.2, págs. 72-85.

Zipp, John F. (2007). "Learning by Exams: The Impact of Two-Stage Cooperative Tests". En: *Teaching Sociology* 35.1, págs. 62-76.