WPOM

Working Papers on Operations Management.
Vol. 8, SpIssue (1-12)
ISSN: 1989-9068

http://dx.doi.org/10.4995/wpom.v8i0.7242

Comparación del efecto de diferentes modos de agregar las calificaciones de evaluación continua en la nota final

Comparison of different ways of computing grades in continuous assessment into the final grade

Marin-Garcia, Juan A.[a], Maheut, Julien[b], Garcia-Sabater, Julio J.[c]

[a,b,c] ROGLE. Dpto. de Organización de Empresas. Universitat Politècnica de València. Camino de Vera S/N 46021 Valencia. [a] jamarin@omp.upv.es, [b] julien.maheut@gmail.com and [c] jugarsa@omp.upv.es.

Recibido: 2017-02-07 Aceptado: 2017-02-08

Abstract

We present the results of comparing various ways of calculating students' final grades from continuous assessment grades. Traditionally the weighted arithmetic mean has been used and we compare this method with other alternatives: arithmetic mean, geometric mean, harmonic mean and multiplication of the percentage of overcoming of each activity. Our objective is to verify, if any of the alternative methods, agree with the student's performance proposed by the teacher of the subject, further discriminating the grade between high and low learning outcomes and reducing the number of approved opportunists.

Keywords: *marks; grade; student performance; higher education; summative assessment; classroom learning*

Objetivo

Nuestro objetivo es comparar los efectos de varias formas de calcular la nota final de los alumnos de una asignatura universitaria, a partir de las calificaciones provenientes de la evaluación continua formativa.

El sistema de evaluación de una asignatura es un aspecto importante porque no sólo condiciona u orienta el modo de trabajar de los estudiantes (Black, Paul & Wiliam, Dylan, 1998; Black, P. & Wiliam, D., 1998; Watts, García-Carbonell, & Llorens, 2006), también tiene un claro efecto en la satisfacción de los estudiantes (Gatfield, 1999; Gibbs & Taylor, 2016; Marin-Garcia, Martínez-Gómez, & Giraldo-O'Meara, 2014; Trotter, 2006; Viles Diez, Zárraga-Rodríguez, & Jaca García, 2013; Walker & Palmer, 2011) y ésta a su vez afecta a la relación del estudiante con la institución, lo que a la larga configura la imagen y el atractivo de las universidades para los estudiantes(Trullas & Enache, 2011).

El problema que se plantea en la asignatura objeto de estudio no es un caso aislado. En diversas reuniones de la Comisión Académica del Título y en conversaciones con diversos profesores del Centro se ha manifestado con excesiva frecuencia el mismo problema. Cuando en una asignatura concurren todas estas situaciones simultáneamente:

- Se propone una evaluación continua donde se realizan muchos ejercicios, o tareas, por parte de los estudiantes
- Algunas (o bastantes) de esas actividades o ejercicios se realizan en grupo y la calificación se comparte entre todos los integrantes del grupo pues es difícil identificar la participación/implicación real de cada persona
- Es sencillo superar cada una de las actividades porque se trata de actividades de corta duración, muy estructuradas, con criterios explícitos de rendimiento (muchas veces soportados por rúbricas que actúan como listas de comprobación) y que cuentan con el apoyo del profesor o de sus compañeros para resolver dudas

Si se incorpora la suma (o promedio) de estas calificaciones a la nota final, es muy sencillo para los alumnos lograr el aprobado, a pesar de que en los actos de evaluación exigentes no demuestren una superación clara de los objetivos de aprendizaje de la asignatura. Estos actos de evaluación exigentes requieren de una elaboración profunda de las respuestas a problemas no sencillos o rutinarios, sin la posibilidad de ayuda o guía por parte del profesor u otros estudiantes y con el tiempo limitado para generar la respuesta. Estos actos, normalmente, se realizan durante los exámenes finales o parciales de la asignatura.

Además de sospechar que los alumnos que demuestran bajo rendimiento aprueban con relativa facilidad la asignatura, también se observa que las notas de los alumnos están concentradas en un rango muy estrecho de nota, sin demasiada discriminación. Pero, al mismo tiempo se comprueba que no todos los alumnos saben lo mismo, ni saben hacer las mismas cosas, ni tienen la misma actitud.

Trabajos relacionados

Hemos realizado una búsqueda de literatura para identificar artículos científicos que hayan tratado problemáticas similares y pueda proponer opciones para resolverla. La estrategia de búsqueda utilizaba estos parámetros:

- Búsqueda en título, resumen y palabras clave: ("final mark" AND course)
- Refined by: DOCUMENT TYPES: (ARTICLE)
- Timespan: All years.

En total se han localizado 21 en WOS y 62 en Scopus, que han resultado en un conjunto de 77 referencias únicas.

En general, los estudios sobre evaluación publicados en revistas científicas se han centrado en analizar cómo conseguir evaluaciones más justas, válidas o fiables (Dalziel, 1998); el uso de estudiantes en el proceso de evaluación (Marin-Garcia, Aragonés Beltran, & Melón, 2014; Potgieter, Ackermann, & Fletcher, 2010; Tejeiro et al., 2012); el impacto del uso de rúbricas en las notas otorgadas a los resultados de ejercicios, o al proceso seguido para completarlo (Marin-Garcia & Santandreu-Mascarell, 2015; Panadero & Jonsson, 2013); el método de evaluación más adecuado en función del objeto a evaluar o para diferentes estrategias metodológicas, entre las que destacan las que incorporan el uso de tecnología de la información en el aula (Bliuc, Ellis, Goodyear, & Piggott, 2010; Bliuc, Ellis, Goodyear, & Piggott, 2011; González-Marcos, Alba-Elías, Navaridas-Nalda, & Ordieres-Meré, 2016; Green, Farchione, Hughes, & Chan, 2014; Sanna, Lamberti, Paravati, & Demartini, 2012) o trabajo en equipo (Perello-Marin, Vidal-Carreras, & Marin-Garcia, 2016; Pratten, Merrick, & Burr, 2014) o aprendizaje basado en problemas (Perez-Benedito,

Perez-Alvarez, & Casati, 2015; Valle, Gonzalvo, & Abril, 2011). Una abrumadora cantidad de estos estudios publicados se ha realizado en disciplinas diferentes de la de Gestión en general o de Gestión de Operaciones en particular (Medina-López, Alfalla-Luque, & Marin-Garcia, 2011).

Sin embargo, hemos encontrado poca investigación relacionada con el modo en que se computa la nota final, a partir de las notas de diferentes actos de evaluación, y cómo esto puedo afectar a la validez, fiabilidad o capacidad de discriminación entre buenos y malos rendimientos de los estudiantes. Quizás una de las pocas excepciones sean: (Dalziel, 1998; Knight, 2002; Knight & Banks, 2003; Yorke, 1998; Yorke, 2010, 2011).

Metodología

Los datos utilizados provienen del curso 2016-17 en una asignatura de master con 21 alumnos. La asignatura se imparte durante 13 semanas lectivas en sesiones de 3,5 horas (una sesión semanal).

Antes de realizar los cálculos de la nota final. El profesor de la asignatura ha calificado a los alumnos siguiendo dos métodos:

- SE0.1: calificación de 0 a 10 holística, basada en los resultados de aprendizaje superados por la persona, utilizando las anotaciones tomadas en observaciones en clase, las preguntas directas a los alumnos, las preguntas que formulan los alumnos y los errores conceptuales que manifiestan. Cada semana el profesor dedicaba unas dos horas y media de clase a esta actividad mientras los alumnos realizaban actividades en su presencia.
- SE0.2: utilizando un método de comparación pareada (Marin-Garcia, Garcia-Sabater, Morant Llorca, & Conejero, 2016) derivada de trabajos previos (Marin-Garcia, Aragonés Beltran, et al., 2014; Marin-Garcia, Ramirez Bayarri, & Atares-Huerta, 2015). El proceso se inicia comparando parejas de alumnos indicando quién, a juicio del profesor, ha tenido más rendimiento respecto a los objetivos de aprendizaje de la asignatura considerados de manera holística. Las puntuaciones tenían 3 categorías (la persona A tiene mayor rendimiento que B, igual rendimiento, la persona B tiene mayor rendimiento que A). El sistema utiliza un algoritmo de grafos para decidir cuáles son las parejas a mostrar. En lugar de ser necesarias las 210 comparaciones posibles, el procedimiento ha convergido a una solución tras 50 comparaciones. Ver Tabla 1. El tiempo invertido por el profesor han sido 10 minutos en total. Una vez identificados los nodos (grupos de alumnos con similares resultados de aprendizaje) se ha procedido a volver a re-escalar las distancias en una nota de 0 a 10. Para ello, se han seleccionado el mejor y el peor nodo, asignándoles una calificación numérica (ver Tabla 2). En este caso el mejor nodo representaba un 10 (MH) y el peor se ha considerado equivalente a una nota de 3. Además, se ha seleccionado un nodo intermedio que represente el punto de corte de "notable" (una nota de 7/10). Al resto de nodos se les ha asignado una nota numérica proporcional a la distancia con los nodos ya calificados.

Entre la calificación SE0.1 y SE0.2 han transcurrido 15 días de intenso trabajo en otras actividades, para garantizar que ambas notas son independientes.

Tabla 1 . Nodos del paso 1 en SE0.2

Grupo	Nombre
10	Al04
9	Al05 Al13
8	Al06 Al07 Al01
7	Al19 Al02
6	Al21
5	Al14 Al17
4	Al11 Al12
3	Al09 Al08 Al18 Al17
2	Al10 Al03 Al15
1	Al20

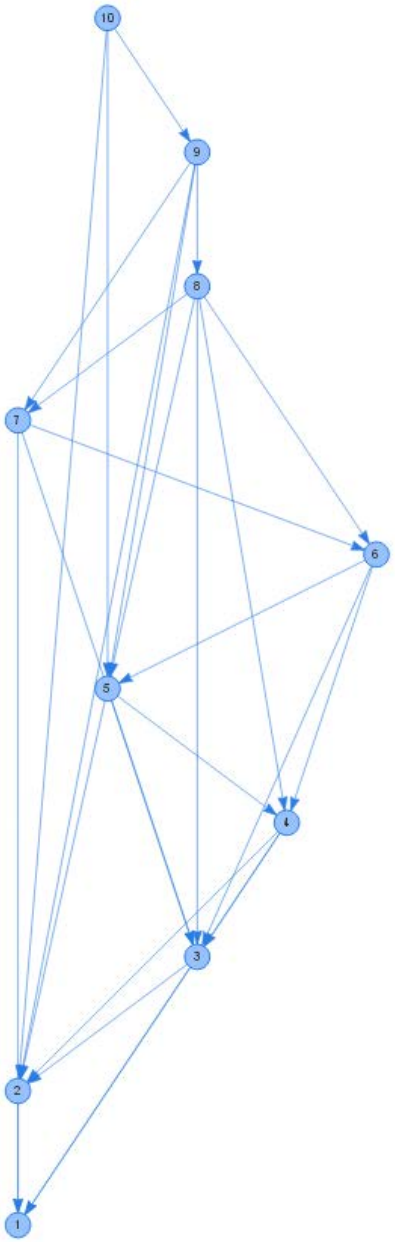
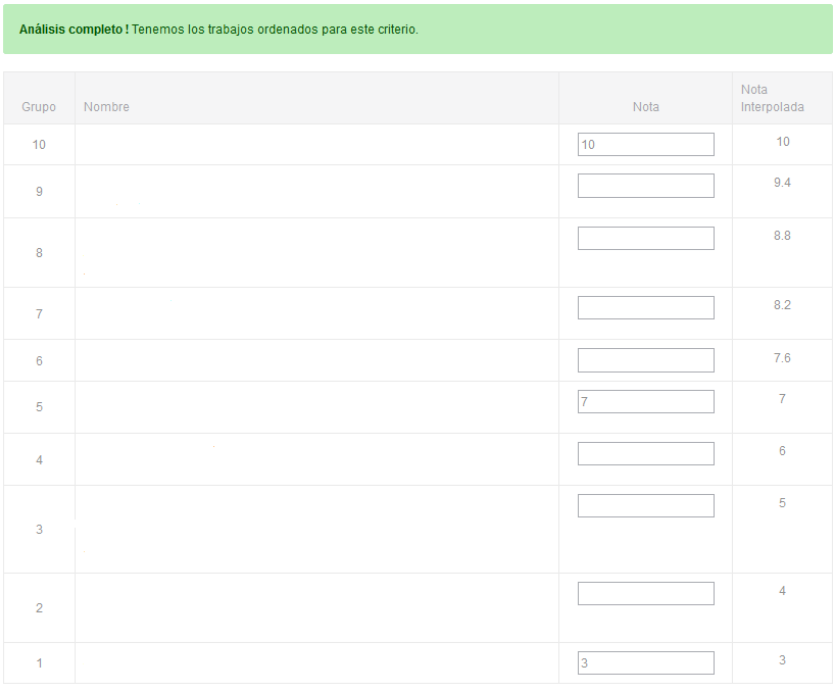


Tabla 2 . Re-escalado de notas. Paso 2 en SE0.2



El sistema de evaluación "oficial" (SE1) es una media ponderada. Consta de 30 registros por cada alumno. Estos registros están agrupados (se calcula el promedio) en 6 categorías cuya ponderación para el cálculo de la nota final de la asignatura aparece reflejado en la Tabla 3. Se han elaborado rúbricas detalladas para puntuar cada una de las actividades que se contemplan en los registros. En algunos casos son listas de comprobación Sí/no y en otros casos son rúbricas analíticas con ejemplos de comportamientos o resultados en 3-5 niveles de desempeño. Los alumnos han dispuesto de una copia de las rúbricas desde el principio del curso y han sido explicadas con detalle antes del inicio de cada una de las actividades. En algunos casos se han utilizado con ejercicios de ejemplos para que los alumnos tuvieran información de cómo se emplearían en la calificación de sus trabajos. El desempeño excepcional en alguno de los trabajos daba origen a notas por encima de 10/10.

Tabla 3 . Componentes del sistema de evaluación oficial de la asignatura

Categoría	Nº de registros	Nota media de la clase	**Peso (%)**
Asistencia (S)	13	8.5	24%
Diario y tareas en el Foro (V)	2	6.9	20%
Exámenes On-line (Y)	3	9.4	17%
Producto individual (T)	3	9.0	18%
Proyecto de mejora en equipo -Proceso de grupo (X)	2	10.0	14%
Puntos profesor (Z)	1	8.7	7%

La nota media obtenida por los alumnos en cada una de las categorías se sitúa entre 6.9 y 10.

Nuestro objetivo es comparar el SE1 con otros modos de calcular la nota y ver cuáles de ellos se ajusta más a las notas globales proporcionadas por el profesor de la asignatura (SE0.1 y SE0.2). Consideraremos el ajuste como el coeficiente de correlación de rangos de Spearman y el coeficiente de correlación intra clase (ICC(2,1) *agreement*) (Hair, Anderson, Tatham, & Black, 1995; Losilla, Navarro, Palmer, Rodrigo, & Ato, 2005; Marin-Garcia, 2009; Marin-Garcia, 2017; Marin-Garcia, Aragonés Beltran, et al., 2014).

Los métodos alternativos considerados son:

- SE2: media aritmética simple de las 6 categorías de evaluación
- SE3: media geométrica de las 6 categorías de evaluación
- SE4: producto del porcentaje de superación de cada de las 6 categorías de evaluación. Se calcula de un modo análogo a como se calcularía el *First Time Through* en una línea de fabricación. Cada componente de la nota se valora de 0% a 100% de resultados de aprendizaje logrado (se puede incluso permitir un 110% o un 120% para resultados que exceden los objetivos de aprendizaje). Luego se multiplican todos esos porcentajes para sacar la nota final (de 0 a 100), y se re-escala de 0 a 10
- SE5: media armónica de las 6 categorías de evaluación

Resultados

En la Tabla 4 resumimos los resultados de las calificaciones finales. Como profesor responsable de la asignatura, tras haber observado sistemáticamente a los alumnos durante las 13 semanas lectivas, la columna que mejor representa los resultados de aprendizaje de los alumnos es la SE0.2.

Tabla 4 . Calificaciones finales de las 21 personas matriculadas, dependiendo del sistema de cómputo elegido

PER-SONA	SE0.1	SE0.2	SE1	SE2	SE3	SE4	SE5
AL01	8	8.8	9.1	9.5	9.3	6.3	9.0
AL02	9.5	8.2	9.3	9.5	9.4	6.7	9.2
AL03	4	4	8.3	8.2	8.0	2.6	7.8
AL04	10	10	10.8	11.0	10.9	17.1	10.9
AL05	9.8	9.4	9.6	10.0	9.9	9.3	9.8
AL06	9.5	8.8	9.6	9.6	9.6	7.7	9.6
AL07	9.5	8.8	10.1	10.1	10.1	10.5	10.1
AL08	6.5	5	7.9	8.3	8.1	2.9	7.9
AL09	6	5	7.8	8.0	7.8	2.3	7.6
AL10	4	4	8.1	7.5	6.1	0.5	3.8
AL11	7	6	7.9	8.1	7.8	2.3	7.5
AL12	6	6	8.2	8.5	8.1	2.8	7.6
AL13	8.5	9.4	9.1	9.3	9.2	6.1	9.0
AL14	8	7	9.4	9.6	9.5	7.6	9.5
AL15	4	4	8.0	8.0	7.7	2.1	7.4
AL16	7	7	7.1	7.5	6.9	1.0	6.2
AL17	5	5	7.8	8.0	7.8	2.3	7.7
AL18	5	5	8.0	8.3	8.1	2.7	7.9
AL19	9	8.2	9.6	9.7	9.7	8.4	9.7
AL20	4	3	7.6	7.2	6.6	0.8	5.9
AL21	7	7.6	10.0	10.2	10.2	11.1	10.1

Las correlaciones de rangos entre los diferentes métodos son todas elevadas (Tabla 5). Esto indica que, independientemente del valor absoluto de la calificación, todos los métodos ordenan de manera bastante similar a los estudiantes. Sin embargo, nosotros estamos interesados no en la correlación, sino en el acuerdo entre puntuaciones, ya que es la puntuación absoluta la que genera la categoría de calificaciones en el acta de la asignatura y queremos ver qué método genera unas calificaciones similares a las del evaluador experto.

Observando los valores de ICC(2,1) podemos comprobar que el método SE0.1 tienen un grado de acuerdo excelente. Es decir, el juicio del experto tiene una fiabilidad elevada tanto al cambiar el método para realizar la calificación, como al dejar pasar el tiempo y hacer un Test/re-test. Si tomamos las calificaciones SE0.2 como "*gold standard*", se manifiesta un elevado desacuerdo con las calificaciones obtenidas con la media aritmética ponderada (SE1) y con la media aritmética simple (SE2). La media geométrica mejora un poco el grado de acuerdo, pero se sigue manteniendo a unos niveles muy bajos. La media armónica presenta unos valores de acuerdo moderados, con la ventaja adicional de que es muy sencilla de calcular por estar implementada en cualquier hoja de cálculo. Por último, el mejor grado de acuerdo se produce con el método de multiplicación de porcentajes (SE4).

Tabla 5 . Grado de acuerdo entre las medidas. En la diagonal inferior correlación Spearman's rho. En la submatriz superior (negrita) las ICC(2,1) de cada variable comparada con SE0.2. ** nivel de significación 1%

	SE0.1	SE0.2	SE1	SE2	SE3	SE4	SE5
SE0.1	-	**0.940**					
SE0.2	0.950**	-	**0.354**	**0.380**	**0.473**	**0.619**	**0.570**
SE1	0.727**	0.733**	-				
SE2	0.811**	0.810**	0.938**	-			
SE3	0.821**	0.811**	0.913**	0.993**	-		
SE4	0.825**	0.810**	0.909**	0.992**	0.999**	-	
SE5	0.809**	0.792**	0.894**	0.970**	0.987**	0.986**	-

Contribución

Presentamos los resultados de comparar varios modos de agregar las calificaciones de evaluación continua para calcular las notas finales de los alumnos. Tradicionalmente se ha usado la media aritmética ponderada, y comparamos ése método con otras alternativas: la media aritmética, la media geométrica, la media armónica y la multiplicación del porcentaje de superación de cada actividad. Nuestro objetivo es comprobar, si alguno de los métodos alternativos, concuerda con el rendimiento de alumnos propuesto por el profesor de la asignatura, discriminando más la nota entre altos y bajos rendimientos y reduciendo el número de aprobados oportunistas

La principal contribución para los profesores es doble. Por un lado, en grupos poco numerosos donde se puede llegar a conocer con detalle el grado de aprendizaje de los alumnos, el método de comparación pareada es el que presenta unas calificaciones más acordes con el juicio experto del profesor. Es curioso que esas calificaciones representan mejor su punto de vista que la puntuación analítica de los trabajos o la puntuación holística sobre una escala de 0 a 10 (aunque este último caso tiene un elevado grado de acuerdo con las calificaciones de comparación pareada). La comparación pareada se hace de una manera muy eficiente en tiempo. Requiriendo solo 10 minutos adicionales respecto a las horas dedicadas a observación y *feedback* formativo realizado durante el curso. La comparación pareada es un método muy poco usado en la evaluación de estudiantes y convendría seguir investigando sobre las ventajas e inconvenientes de este método, que parece prometedor, con los datos del caso estudiado en este artículo.

Por otro lado, en grupos masificados no es viable la comparación pareada, pues es difícil tener un conocimiento del aprendizaje real de los alumnos. Sin embargo, si pudiéramos generalizar los datos de este trabajo, el uso de la media armónica o la multiplicación de porcentajes de rendimiento de las actividades, podrían ser un sustitutivo bastante más acertado que la media aritmética ponderada de múltiples actos de evaluación para emular el juicio de experto en la evaluación de alumnos.

Ambas contribuciones profesionales representan también retos interesantes para la investigación sobre la evaluación del aprendizaje en las universidades, abriendo unas líneas de trabajo poco exploradas hasta el momento.

También es posible que estas líneas de investigación se puedan extender al área de evaluación del desempeño en gestión de Recursos Humanos, donde es posible encontrar problemas para poder discernir entre los altos y bajos rendimientos.

Este trabajo no está exento de limitaciones. La principal es el escaso número de datos manejados para realizar la comparación de métodos. Sería recomendable ampliar el análisis a datos de otros cursos académicos de asignaturas similares (máster con menos de 30 alumnos por grupo) y de asignaturas con grupos más números, de otros niveles (grado con 70-90 alumnos por grupo) y de otras titulaciones.

Por otra parte, ninguno de los métodos con calificaciones analíticas (SE1 a SE5) supera la limitación de asumir propiedades numéricas a unas simples clasificaciones (Dalziel, 1998), por lo que sería interesante plantear, como investigación futura, la posibilidad de usar "*concept mapping*" (Kane & Trochim, 2007) como herramienta para integrar diferentes calificaciones.

También sería interesante comparar los resultados que saldrían al emplear una media geométrica ponderada.

Por último, queda pendiente analizar el efecto de notas extremadamente bajas en las medias armónicas o en la multiplicación de porcentajes. La media armónica no se puede calcular si alguno de los componentes es cero. Se podría resolver sustituyendo los ceros por un valor bajo (0.001 ó 1) antes del cómputo de la media. Sin embargo, esta media da más peso a los valores pequeños, de modo que hay que analizar el efecto que produce, en el resultado final, un valor extremadamente pequeño. Respecto a la multiplicación de porcentajes (SE4), este método convierte todos los actos en obligatorios. Una opción a valorar es que varias calificaciones se agrupen en una categoría (pudiendo elegir para agruparlos la media, la media descartando los peores, o cualquier otra opción). En ése casos, la multiplicación de porcentajes se hace sobre las calificaciones 0% a 100% de las categorías.

Agradecimientos

Este trabajo ha sido parcialmente financiado por la Universitat Politécnica de Valencia (PIME/2016/A/027/A) "La evaluación pareada como metodología para la evaluación del pensamiento crítico de los alumnos".

Referencias

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. doi:10.1080/0969595980050102

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Bliuc, A. M., Ellis, R., Goodyear, P., & Piggott, L. (2010). Learning through face-to-face and online discussions: Associations between students' conceptions, approaches and academic performance in political science. *British Journal of Educational Technology, 41*(3), 512-524. doi:10.1111/j.1467-8535.2009.00966.x

Bliuc, A. M., Ellis, R. A., Goodyear, P., & Piggott, L. (2011). A blended learning approach to teaching foreign policy: Student experiences of learning through face-to-face and online discussion and their relationship to academic performance. *Computers and Education, 56*(3), 856-864. doi:10.1016/j.compedu.2010.10.027

Dalziel, J. (1998). Using marks to assess student performance, some problems and alternatives. *Assessment and Evaluation in Higher Education, 23*(4), 351-366. doi:10.1080/0260293980230403

Gatfield, T. (1999). Examining student satisfaction with group projects and peer assessment. *Assessment & Evaluation in Higher Education, 24*(4), 365-377.

Gibbs, J. C., & Taylor, J. D. (2016). Comparing student self-assessment to individualized instructor feedback. *Active Learning in Higher Education, 17*(2), 111-123. doi:10.1177/1469787416637466

González-Marcos, A., Alba-Elías, F., Navaridas-Nalda, F., & Ordieres-Meré, J. (2016). Student evaluation of a virtual experience for project management learning: An empirical study for learning improvement. *Computers and Education, 102*, 172-187. doi:10.1016/j.compedu.2016.08.005

Green, R. A., Farchione, D., Hughes, D. L., & Chan, S. P. (2014). Participation in asynchronous online discussion forums does improve student learning of gross anatomy. *Anatomical Sciences Education, 7*(1), 71-76. doi:10.1002/ase.1376

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis* (4º ed.). New Jersey: Prentice Hall.

Kane, M., & Trochim, W. M. K. (2007). *Concept mapping for planning and evaluation* (Vol. 50). London: SAGE.

Knight, P. T. (2002). Summative assessment in higher education: Practices in disarray. *Studies in Higher Education, 27*(3), 275-286. doi:10.1080/03075070220000662

Knight, P. T., & Banks, W. M. (2003). The assessment of complex learning outcomes. *Global Journal of Engineering Education, 7*(1), 39-49.

Losilla, J. M., Navarro, J. B., Palmer, A., Rodrigo, M. F., & Ato, M. (2005). *Análisis de datos. Del contraste de hipótesis al modelado estadístico*. Barcelona: Edicions a Petició.

Marin-Garcia, J. A. (2009). Los alumnos y los profesores como evaluadores. Aplicación a la calificación de presentaciones orales. *Revista Espanola De Pedagogia, 67*(242), 79-97.

Marin-Garcia, J. A. (2017). Protocol: Inter-rater and intra-rater consistency validation of a rubric to assess oral presentation skills for university students. *WPOM-Working Papers on Operations Management, 7*(2), (in press).

Marin-Garcia, J. A., Aragonés Beltran, P., & Melón, G. (2014). Intra-rater and inter-rater consistency of pair wise comparison in evaluating the innovation competency for university students. *WPOM-Working Papers on Operations Management, 5*(2), 24-46. doi:http://dx.doi.org/10.4995/wpom.v5i2.3220

Marin-Garcia, J. A., Garcia-Sabater, J. P., Morant Llorca, J., & Conejero, J. A. (2016). *Passam: Peer assessment and monitoring system.* Paper presented at the Congreso Nacional de Innovación Educativa y Docencia en Red- Universitat Politècnica de València-Valencia 07/07/16 al 08/07/16.

Marin-Garcia, J. A., Martínez-Gómez, M., & Giraldo-O'Meara, M. (2014). Redesigning work in university classrooms: Factors related to satisfaction in engineering and business administration students. *Intangible Capital, 10*(5), 1026-1051.

Marin-Garcia, J. A., Ramirez Bayarri, L., & Atares-Huerta, L. (2015). Protocol: Comparing advantages and disadvantages of rating scales, behavior observation scales and paired comparison scales for behavior assessment of competencies in workers. A systematic literature review. *WPOM-Working Papers on Operations Management, 2*(6), 49-63. doi:http://dx.doi.org/10.4995/wpom.v6i2.4032

Marin-Garcia, J. A., & Santandreu-Mascarell, C. (2015). What do we know about rubrics used in higher education? *Intangible Capital, 11*(1), 118-145. doi:http://dx.doi.org/10.3926/ic.

Medina-López, C., Alfalla-Luque, R., & Marin-Garcia, J. A. (2011). Research in operations management teaching: Trends and challenges. *Intangible Capital, 7*(2), 507-548.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*(0), 129-144.

Perello-Marin, M. R., Vidal-Carreras, P. I., & Marin-Garcia, J. A. (2016). What do undergraduates perceive about teamwork? *International Journal of Engineering Education, 32*(3), 1171-1181.

Perez-Benedito, J. L., Perez-Alvarez, J., & Casati, M. J. (2015). Pbl in the teaching of design in aeronautical engineering: Application and evolution of a consolidated methodology. *International Journal of Engineering Education, 31*(1), 199-208.

Potgieter, M., Ackermann, M., & Fletcher, L. (2010). Inaccuracy of self-evaluation as additional variable for prediction of students at risk of failing first-year chemistry. *Chemistry Education Research and Practice, 11*(1), 17-24. doi:10.1039/c001042c

Pratten, M. K., Merrick, D., & Burr, S. A. (2014). Group in- course assessment promotes cooperative learning and increases performance. *Anatomical Sciences Education, 7*(3), 224-233. doi:10.1002/ase.1397

Sanna, A., Lamberti, F., Paravati, G., & Demartini, C. (2012). Automatic assessment of 3d modeling exams. *IEEE Transactions on Learning Technologies, 5*(1), 2-10. doi:10.1109/tlt.2011.4

Tejeiro, R. A., Gómez-Vallecillo, J. L., Romero, A. F., Pelegrina, M., Wallace, A., & Emberley, E. (2012). Summative self-assessment in higher education: Implications of its counting towards the final mark. *Electronic Journal of Research in Educational Psychology, 10*(2), 789-812.

Trotter, E. (2006). Student perceptions of continuous summative assessment. *Assessment & Evaluation in Higher Education, 31*(5), 505-521. doi:10.1080/02602930600679506

Trullas, I., & Enache, M. (2011). Theoretical analysis of the antecedents and the consequences of students' identification with their university and their perception of quality. *Intangible Capital, 7*(1), 170-212. doi:10.3926/ic.2011.v7n1.p170-212

Valle, A. R. A., Gonzalvo, M. J. M., & Abril, F. S. (2011). Is there an alternative to master classes? An ocular physiology experience as part of an optics and optometry degree course. *Arbor, 187*(EXTRA 3), 189-194. doi:10.3989/arbor.2011.Extra-3n3143

Viles Diez, E., Zárraga-Rodríguez, M., & Jaca García, C. (2013). Tool to assess teamwork performance in higher education. *Intangible Capital; Vol 9, No 1 (2013)DO - 10.3926/ic.399*.

Walker, D. J., & Palmer, E. (2011). The relationship between student understanding, satisfaction and performance in an australian engineering programme. *Assessment and Evaluation in Higher Education, 36*(2), 157-170. doi:10.1080/02602930903221451

Watts, F., García-Carbonell, A., & Llorens, J. (2006). Introducción a la evaluación compartida: Investigación multidisciplinar. In F. Watts & A. García-Carbonell (Eds.), *La evaluación compartida: Investigación multidisciplinar* (1 ed., pp. 1-9). Valencia: Editorial de la UPV

Yorke, M. (1998). The management of assessment in higher education. *Assessment & Evaluation in Higher Education*, *23*(2), 101-116.

Yorke, M. (2010). How finely grained does summative assessment need to be? *Studies in Higher Education*, *35*(6), 677-689. doi:10.1080/03075070903243118

Yorke, M. (2011). Summative assessment: Dealing with the 'measurement fallacy'. *Studies in Higher Education*, *36*(3), 251-573. doi:10.1080/03075070903545082