WPOM

Working Papers on Operations Management.
Vol. 5, Nº2 (24-46)
ISSN: 1989-9068

http://dx.doi.org/10.4995/wpom.v5i2.3220

Intra-rater and inter-rater consistency of pair wise comparison in evaluating the innovation competency for university students

Consistencia intra e inter evaluador de la comparación pareada en la evaluación de la competencia de innovación de estudiantes universitarios[1]

Juan A. Marin-Garcia[a], Pablo Aragonés Beltrán[b] y Mónica García Melón[b]

[a] ROGLE. Dpto. de Organización de Empresas. Universitat Politècnica de València. Camino de Vera S/N 46021 Valencia. jamarin@omp.upv.es., [b] Grupo de Valoración y decisión multicriterio. Departamento de Proyectos de Ingeniería, Universidad Politécnia de Valencia.

Recibido: 2014-07-28 Aceptado: 2014-12-20

Abstract

The aim of this paper is to propose a critical example to assess student competencies using the paired comparison as multiple criteria decision making tool based on analytic hierarchy process (AHP). We apply it to innovation competency in a subject with few students (10 master's students, divided into two groups). With the proposed methodology it is possible to create an ordered list for each group, having a distance between students (in one dimension) that reflects the degree of learning (knowledge, skills and attitudes). We found that the scores given by an evaluator for a student are consistent with the rest of the evaluation scores given to the other students in the group. It has also been found that the scores obtained by various sources (teacher, student self-assessment and evaluation of student peers) are consistent with each other. In the future, we should check whether this model is viable with large groups and we should propose a variant that allows sorting, in a list, people from different groups or convert the relative distance between group members in an absolute measure of the degree of achieving the learning outcomes set for the course.

Keywords: *AHP; higher education; student assessment; peer assessment; self assessment; teacher assessment; multicriteria decision method.*

Resumen

El objetivo de este artículo es proponer un ejemplo crítico del uso de la comparación pareada basada en el proceso de decisión multicriterio de jerarquización analítica (AHP) para evaluar competencias de los alumnos. Lo aplicaremos a la competencia de innovación en una asignatura con pocos alumnos (10 alumnos de máster, repartidos en dos grupos). Con la metodología propuesta es posible crear, en cada grupo, una lista ordenada desde la persona que más domina una competencia a la que menos la domina, teniendo una distancia entre ellos (en una dimensión) que refleje el grado de aprendizaje (conocimientos, habilidades y actitudes) relativo entre las personas del grupo. Hemos comprobado que las puntuaciones que da un evaluador para un alumno son consistentes con el resto de puntua-

[1] Una versión anterior, más limitada y reducida, de esta investigación ha sido presentada en los congresos: AcedeDOT-OMTech V Workshop in Operations Management and Technology, Pamplona, 3-4 abril 2014; e IN-RED 2014- Jornadas de Innovación Educativa y Docencia en Red. UPVal. 15-16 julio 2014.

ciones que da ese evaluador al resto de alumnos del grupo. También se ha comprobado que las puntuaciones obtenidas por varias fuentes (profesor, auto-evaluación del alumno y evaluación del alumno a sus compañeros) son consistentes entre sí. En el futuro, deberíamos comprobar si es viable este modelo con grupos numerosos y proponer una variante que permita ordenar, en una misma lista, a las personas de diferentes grupos o convertir la distancia relativa entre las personas del grupo en una media absoluta de grado de consecución de los resultados de aprendizaje planteados para la asignatura

Palabras clave: *AHP; Evaluación de estudiantes; educación universitaria; evaluación por compañeros; auto-evaluación; evaluación del profesor; método de decisión multicriterio.*

Introducción

Desde hace un tiempo, existe un creciente interés por enfocar las titulaciones universitarias hacia el desarrollo de competencias profesionales (Lohmann et al., 2006). Para ello, es necesario no sólo identificar cuáles son las competencias prioritarias para el perfil profesional de la titulación, sino también, los resultados de aprendizaje; los niveles que discriminan entre un resultado pobre y uno sobresaliente; los instrumentos para evaluar la adquisición de la competencia y el modo en que vamos a proporcionar información a los alumnos para que puedan progresar en su aprendizaje (Marin-Garcia et al., 2009a; Terry et al., 2002).

Una de las competencias más demandadas en las sociedades avanzadas es la competencia de innovación, que aparece reflejada en la mayoría de los planes de estudios de España bajo diferentes nombres (innovación, creatividad, emprendimiento, habilidad para tomar decisiones, capacidad para resolver problemas...) (Andreu Andrés y García-Casas, 2014; Ingols y Shapiro, 2014; Marin-Garcia et al., 2008; Mula et al., 2012).

La evaluación de competencias es una tarea compleja que incluye decidir sobre varios aspectos o dimensiones de la competencia. En general, par la mayoría de las competencias profesionales, no suele haber demasiado acuerdo acerca de qué subdimensiones (capacidades o habilidades) las componen y, por lo tanto, qué habilidades deben evaluarse (Lohmann et al., 2006; Marin-Garcia et al., 2013).

Por otra parte, no siempre es fácil que los profesores puedan o sepan observar y valorar determinadas habilidades. Por ello, se ha recomendado, en ocasiones, incorporar a los alumnos como evaluadores, bien de sus propias habilidades, o bien de las habilidades de sus compañeros con los que interactúan muchas horas a lo largo del curso. Tanto la auto-evaluación como la evaluación de compañeros presenta beneficios pues aumenta la capacidad crítica y de observación de los estudiantes (Van Overveld y Verhoeff, 2013). Del mismo modo, fomenta la reflexión y la meta-cognición, que son dos componentes importantes en el aprendizaje autónomo a lo largo de la vida (Lind et al., 2002; Ljungman y Silén, 2008).

No obstante, a pesar de los beneficios citados, a los estudiantes les supone un reto, a veces insalvable, el asignar una nota usando una puntuación en una escala de 0 a 10 y no todos los alumnos están dispuestos a participar como evaluadores, ni son igual de fiables cuando evalúan (Pond, 2007; Van Overveld y Verhoeff, 2013).

Para superar esos problemas, se ha sugerido el empleo de rúbricas (Andreu Andrés y García-Casas, 2014; Marin-Garcia, 2009), a lo que nosotros añadimos en esta investigación, el utilizar como método de evaluación las comparaciones pareadas (Van Overveld y Verhoeff, 2013).

El objetivo de esta investigación es comprobar, en una asignatura con pocos alumnos, si las notas que ponemos los profesores en el sistema de evaluación de la asignatura ordenan a los alumnos en un orden parecido al contrastar varias fuentes –observaciones del profesor, auto evaluación del alumno y evaluación de los compañeros-. Junto con eso, desarrollamos un método que, en el futuro, deberíamos comprobar si es viable incluso con grupos numerosos. De este modo, podremos extender las conclusiones de estudios actuales que consideran que las puntuaciones de alumnos y profesores no son significativamente diferentes (Andreu Andrés y García-Casas, 2014; Lind et al., 2002; Marin-Garcia, 2009) y aportaremos una nueva forma de captura de datos de evaluación que simplifica la tarea del evaluador, mantiene la precisión de las puntuaciones y exige un tiempo razonable y sostenible para ser completada tanto por alumnos como por profesores.

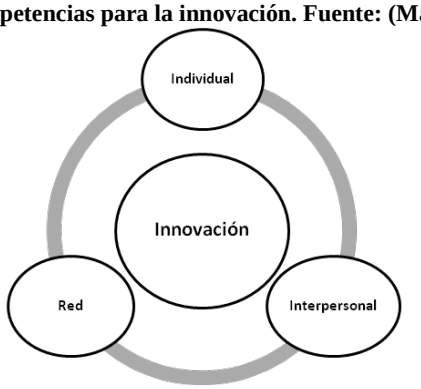
Dimensiones de la competencia de innovación en estudiantes universitarios

La innovación es un proceso que permite la creación, adaptación, implantación o explotación de nuevos productos, servicios, métodos o procedimientos, suministros, mercados, modelo de negocio o modelos de gestión; que permiten la mejora del desempeño de una organización, equipo o persona (Comisión Europea, 1995; Gee, 1981; Goffin y Mitchell, 2010; González Pernía y Peña-Legazkue, 2007; Goswami y Mathew, 2005; Klippel et al., 2008; Lawson y Samson, 2001; Lehto et al., 2011; Lyons et al., 2007; Marin-Garcia et al., 2011; Mol y Birkinshaw, 2009; Schumpeter, 1934; Tonnessen, 2005; Vaccaro et al., 2012). Sin embargo, no está tan claro en qué consiste, como desarrollar o como medir la competencia que permite a una persona ser innovadora o contribuir a la innovación de una organización.

Una competencia se puede definir como el desempeño adecuado de tareas en contextos complejos que requieren la activación e integración de conocimientos, procedimientos de decisión, capacidades, habilidades, actitudes y valores con autonomía y responsabilidad (European Commission, 2008; Fernández March, 2010; Lasnier, 2000; Perrenoud, 2005; Villa Sánchez y Poblete, 2007).

Centrándonos en la competencia de innovación (Figura 1), podemos considerar tres dimensiones en las que agrupar habilidades o comportamientos: individual, interpersonal y red (Lehto et al., 2011; Marin-Garcia et al., 2013; Penttilä y Kairisto-Mertanene, 2012; Watts et al., 2012; Watts et al., 2013).

Figura 1. Modelo de competencias para la innovación. Fuente: (Marin-Garcia et al., 2013)



Comparación pareada como método de evaluación

Podemos considerar tres posibles métodos de recogida de los datos de entrada necesarios para la evaluación de los alumnos. Las listas ordenadas (rankings), las escalas de puntuación (ratings) y la comparación pareada (*paired comparisons*) (Hatzinger y Dittrich, 2012).

En algunas ocasiones, podemos ordenar a los estudiantes o sus trabajos desde el mejor al peor, creando una lista ordenada bajo determinado criterio de rendimiento académico. En otras ocasiones, podemos evaluar a los estudiantes mediante escalas de puntuación. Estas escalas pueden ser de tipo numérico (con una calificación de 0 a 10, por ejemplo, en cada uno de los criterios a evaluar) o escalas tipo Likert que permiten recoger también valoraciones cualitativas como sentimientos, opiniones o actitudes. La ponderación de los valores obtenidos en las escalas de puntuación o de las frecuencias de las alternativas de respuesta categóricas, en las diferentes dimensiones, permite obtener una calificación global, que puede usarse para ordenar a los estudiantes de mejor a peor rendimiento. Por último, existe también la posibilidad de utilizar las comparaciones pareadas. En este tercer caso, para cada criterio de evaluación, se forma una pareja de estudiantes y se valora cual de los estudiantes de esa pareja posee en mayor medida el atributo evaluado y luego se pasa a la siguiente pareja hasta completar todas las parejas posibles que se pueden formar con los estudiantes (Hatzinger y Dittrich, 2012).

Es evidente que una vez usado uno de los tres métodos de captura de datos, se puede trasladar a los otros formatos fácilmente. Es decir, los alumnos con mejor puntuación en una escala, los podemos ordenar en las primeras posiciones de la lista ordenada y deberían ser la opción "ganadora" en la comparación pareada con alumnos con peor puntuación en la misma escala (Hatzinger y Dittrich, 2012). Lo que no es habitual es que se utilicen dos o tres de estos métodos para comprobar si la posición ordenada de los alumnos converge al usar, de manera independiente, diferentes métodos de recogida de datos. Tampoco hemos encontrado demasiados estudios que analicen si la fiabilidad de las puntuaciones o la dificultad del proceso de evaluación cambia al usar uno de estos tres métodos, ni si estas diferencias tienen mayor o menor incidencia en función de si se trata de la evaluación del profesor, la auto-evaluación de los estudiantes o la evaluación de los estudiantes a sus compañeros.

Por otra parte, parece que los estudiantes experimentan dificultades a la hora de usar listas ordenadas o escalas de valoración con sus compañeros (Ikehara y Toyoda, 2012; Pond, 2007; Van Overveld y Verhoeff, 2013). Al mismo tiempo, ciertas experiencias con comparación pareada permiten comprobar la convergencia (fiabilidad *inter-rater* e *intra-rater*) de las puntuaciones de varios evaluadores (Conlon et al., 2012; Yiu et al., 2007) y la validez de las mismas cuando se administra a través de un cuestionario web completado por estudiantes o por profesores como evaluadores (Kan Ma et al., 2013; Marrin et al., 2004).

Estos resultados parece que mejoran en la medida en que se utilizan anclas verbales para las puntuaciones y que se usen entre 7 y 9 niveles en la comparación (De Beuckelaer et al., 2013; Yiu et al., 2007) para proporcionar suficiente rango de variabilidad, sin complicar en exceso el proceso de comparación. Estos criterios coinciden con los planteados por Saaty (1980) a la hora de proponer su escala de comparación binaria y que nosotros hemos adaptado para la comparación del rendimiento de estudiantes (ver Tabla 1)

Tabla 1.- Escala de comparación binaria adaptada de Saaty (1980) para comparación de rendimiento de estudiantes

Escala numérica	Escala verbal	Explicación
1	Iguales	Dos estudiantes tienen habilidades/conocimientos semejantes
3	Moderadamente mejor	La experiencia y la habilidad están a favor de una de las personas
5	Fuertemente mejor	Una persona es fuertemente más competente que la otra en el aspecto evaluado
7	Muy fuertemente mejor	Una persona es mucho más competente que la otra
9	Extremadamente mejor	Una persona está totalmente en un nivel superior a otra
2,4,6,8	Valores intermedios entre dos juicios adyacentes	Se usan como valor intermedio entre las anclas de juicios explícitas de los niveles 1, 3, 5, 7 y 9

Metodología

Aplicación del Proceso de Jerarquización Analítica (AHP) a la evaluación de estudiantes

En esta investigación asumimos que la evaluación es una decisión multicriterio, puesto que para evaluar a un alumno hay que considerar diferentes conocimientos, capacidades o habilidades que no siempre son complementarias. Esto significa que, algunos alumnos tienen un rendimiento bueno o excelente para un aspecto, pero puede ser malo o peor para otro aspecto evaluado.

De los diferentes métodos de decisión multicriterio hemos seleccionad AHP porque el problema de la evaluación de estudiantes puede modelizarse como una jerarquía de objetivos. La complejidad inherente a la decisión de evaluación que contempla múltiples criterios, puede ser resuelta a través de la construcción de estructuras jerárquicas que representan el objetivo, los criterios para evaluar ese objetivo y el modo en que las alternativas evaluadas (alumnos en nuestro caso) se ajustan a esos criterios (Albayrak y Erensal, 2004; Chin et al., 2002; Kasirian et al., 2010; Khalaf y El Mokadem, 2011; Li et al., 2009; Saaty, 1980; Veronese Bentes et al., 2012; Yao, 2010). Una de las principales ventajas del método es que resulta fácil de explicar a las personas que tienen que valorar los diferentes criterios o alternativas, es muy intuitivo y permite trabajar de forma sistemática. Todas estas características son especialmente importantes cuando se incorporan estudiantes al proceso de puntuación (Marin-Garcia, 2009).

Los pasos básicos de los que consta este método se detallan a continuación.

Definir el problema de evaluación/priorización

En nuestro caso queremos ordenar a los alumnos participantes en un grupo en función del grado de rendimiento académico en la competencia de innovación. Los grupos se crearon para elaborar un proyecto de mejora de proceso con la metodología *Kaizen Workshop* (Marin-Garcia et al., 2009b). Los participantes interactuaron durante unas 20 horas presenciales en clase, a lo largo de 8 semanas, realizando diversas actividades relacionadas con la mejora de un proceso para realizar un servicio más eficaz y eficientemente.

Identificar los criterios y las alternativas pertinentes para resolver el problema

De acuerdo con el modelo mostrado en la Figura 1, la competencia de innovación se compone de las capacidades para: ser innovador individualmente (Ind), hacer actuar a otros y potenciar su capacidad individual de innovar por medio de la interacción con el grupo (Grup) y para crear/emplear una red de contactos de modo que permitan al grupo buscar soluciones adecuadas en un entorno más amplio del habitual (Red). Las alternativas a valorar son, en este caso, 10 estudiantes a los que valoraremos el rendimiento en los criterios de decisión seleccionados en el próximo paso.

Estructurar el problema en una jerarquía de diferentes niveles que constituyen: la meta, criterios, subcriterios y alternativas

Cada uno de los alumnos será evaluado usando 26 criterios (Tabla 2). Uno de los criterios (crit00) representa una versión sintética/holística de la competencia global para ser una persona innovadora. El resto de los 25 criterios (crit01 a crit25) son comportamientos asociados a cada una de las tres capacidades en las que hemos desglosado la competencia de innovación (Ind, Grup y Red). Con estos 25 criterios, calcularemos una media ponderada (Pond) que pretende representar una versión analítica de la competencia global para ser una persona innovadora.

Tabla 2.- Dimensiones de la competencia de innovación. Fuente: (Watts et al., 2013)

Dimension	Id	Descripción
Global	crit00	Ser una persona innovadora. Valora la competencia para fomentar la innovación, considerando la innovación como un proceso que permite la creación y puesta en práctica de algo (producto o servicio, métodos, mercados, suministros, modelos de negocios, modelos de gestión o tareas en general) que sea percibido como novedad por la organización/grupo y que persiga una mejora del rendimiento o resultados de la organización/grupo
Ind	crit01	Propongo nuevas maneras de poner en marcha las ideas
Ind	crit02	Anticipo cómo se pueden desarrollar los acontecimientos
Grup	crit03	Colaboro activamente
Grup	crit04	Contribuyo al buen funcionamiento del grupo
Red	crit05	Tengo en cuenta el impacto social de la tarea
Grup	crit06	Hago que los demás contribuyan en las tareas del grupo
Grup	crit07	Escucho a mis compañeros de grupo
Grup	crit08	Afronto los conflictos con flexibilidad con el fin de alcanzar acuerdos
Ind	crit09	Identifico las relaciones entre las diferentes partes o aspectos de la tarea
Ind	crit10	Propongo ideas relacionadas con la resolución de la tarea
Grup	crit11	Soy perseverante en la consecución de los objetivos
Grup	crit12	Oriento la tarea hacia el objetivo
Red	crit13	Uso contactos externos al grupo para que nos ayuden a alcanzar nuestros objetivos
Red	crit14	Puedo trabajar en grupos multidisciplinares
Grup	crit15	Transmito ideas de manera efectiva
Ind	crit16	Evalúo las ventajas y desventajas de las acciones
Grup	crit17	Establezco un ambiente constructivo en el grupo a través del diálogo
Red	crit18	Utilizo hábilmente los recursos disponibles
Red	crit19	Puedo trabajar en grupos multiculturales
Ind	crit20	Abordo la tarea desde diferentes puntos de vista
Ind	crit21	Presento ideas creativas

Grup	crit22	Tomo iniciativas
Grup	crit23	Muestro entusiasmo
Red	crit24	Asumo riesgos audaces pero razonables
Red	crit25	Aplico valores éticos en las decisiones del grupo
Global	Pond	Ponderación de los valores de los criterios 1 a 25

Hacer comparaciones pareada en cada nivel jerárquico con respecto a un elemento específico (sean criterios o alternativas a valorar)

En cada nivel jerárquico las comparaciones pareadas se hacen con sentencias que utilizan valores numéricos tomados de una escala de 1-9 (ver Tabla 1).

Para los criterios, se compara cada criterio i con cada criterio j, respondiendo a la pregunta ¿es el criterio i igual, …, extremadamente más importante que el criterio j? Se construirán 4 matrices rellenadas por el profesor de la asignatura. En la primera (nivel 1) se compararán, por parejas, la importancia de las tres capacidades (Ind, Grup y Red). Las otras tres matrices (nivel 2) son: la comparación pareada de los 7 comportamientos de la capacidad Individual, la comparación pareada de los 11 comportamientos de la capacidad de grupo y, por último la comparación pareada de los 7 comportamientos de la capacidad en red.

Para valorar las alternativas, usando la escala de la Tabla 1, se responde a: respecto al criterio k, ¿es la alternativa i igual, …, extremadamente mejor, que la j ? Se rellenan 26 matrices (una para cada criterio) por cada evaluador. El profesor será evaluador en los dos grupos y cada uno de los estudiantes evaluará a todos los componentes del grupo (incluido a sí mismo). En la Figura 2 mostramos un ejemplo del formulario para la construcción de la matriz de comparación pareada de crit03 en uno de los grupos.

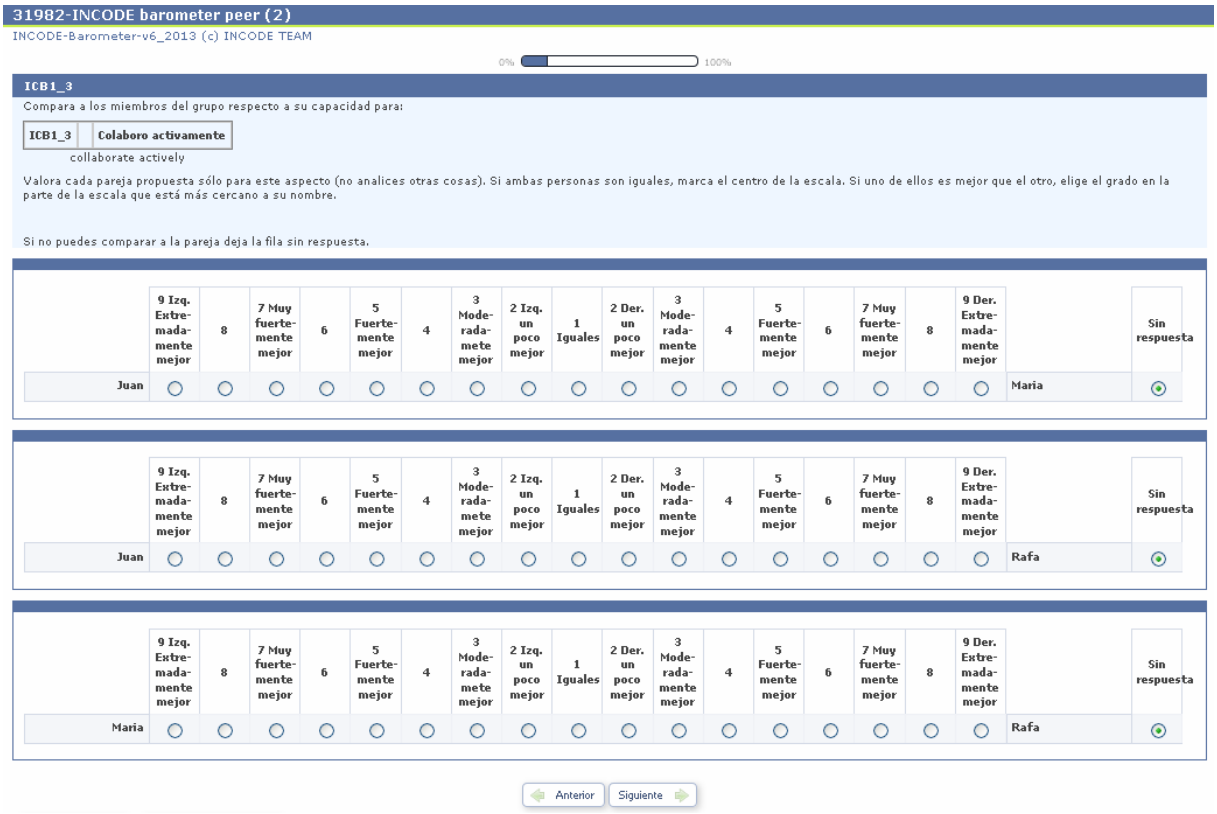
Construir matrices de decisión (A)

Las comparaciones del punto 4, llevan a matrices de decisión que son positivas y recíprocas (aij = 1/ aji). El AHP consigue así tener todos los datos en una sola escala unidimensional de prioridades. La recolección de datos para cada una de las matrices requiere n (n-1)/2 comparaciones, donde n es el número de elementos a comparar. Los elementos de la diagonal (comparación de un elemento consigo mismo) son iguales a '1 ' y, el resto de elementos, al ser una matriz recíproca, serán simplemente el inverso del término correspondiente.

Obtención de las prioridades en cada nivel de la jerarquía.

Una vez se dispone de la matriz de juicios de decisión (A) de comparaciones de los criterios o alternativas, las prioridades de estos elementos, se pueden obtener de manera aproximada con el método de la media geométrica. Se calcula la media geométrica de los n elementos de cada fila de la matriz y se normaliza (B) el vector obtenido (se suman las n medias geométricas y se divide cada media geométrica por el total de la suma de todas). Este método proporciona unos resultados suficientemente cercanos al método exacto y es más sencillo de aplicar que el método exacto.

Figura 2.- Cuestionario para la comparación pareada



Cálculo de la consistencia

Debido a que los juicios son subjetivos, éstos no siempre cumplen a la perfección la propiedad de la transitividad ($r_{ij} \cdot r_{jk} = r_{ik}$) , y por ello se dice que no son consistentes. La consistencia de cada una de las matrices se puede calcular con el Ratio de Consistencia (CR por sus siglas en inglés)

$$CI = \frac{\lambda_{max} - n}{n-1} \qquad CR = \frac{CI}{RI}$$

Donde:

CI es el Índice de Consistencia (*Consistency Index*)

n el número de elementos que se comparan entre sí en la matriz de juicios

λ_{max}: autovalor principal. Se calcula multiplicando la matriz A (matriz de decisión) por el vector B (media geométrica normalizada). Este resultado (C) se divide por B componente a componente (vector D). De modo que $D_i = C_i/B_i$. La media aritmética de los elementos del vector D es el valor aproximado de λ_{max}.

RI el Índice Aleatorio (*Random Index*) que depende del tamaño de la matriz de comparaciones:

n	1	2	3	4	5	6	7	8	9	10
RI	0	0	0,58	0,9	1,12	1,24	1,32	1,41	1,45	1,49

Con n>= 5, si el CR es mayor que el 10% se considera que los juicios introducidos en la matriz no son suficientemente consistentes y, por ello, deberían ser revisados (si n=4 el valor de corte es 8%; y si n=3 el valor de corte es 5%)

Obtención de la puntuación ponderada (U(Alternativa$_i$))

Es la agregación, por suma ponderada con el peso global de cada uno de los criterios, de los pesos (prioridades) de cada alternativa (Figura 4). Para calcular el peso global se multiplican los pesos de la rama del criterio para cada uno de los niveles (Figura 3)

Figura 3.- Calculo del peso global de un criterio a partir de los pesos locales

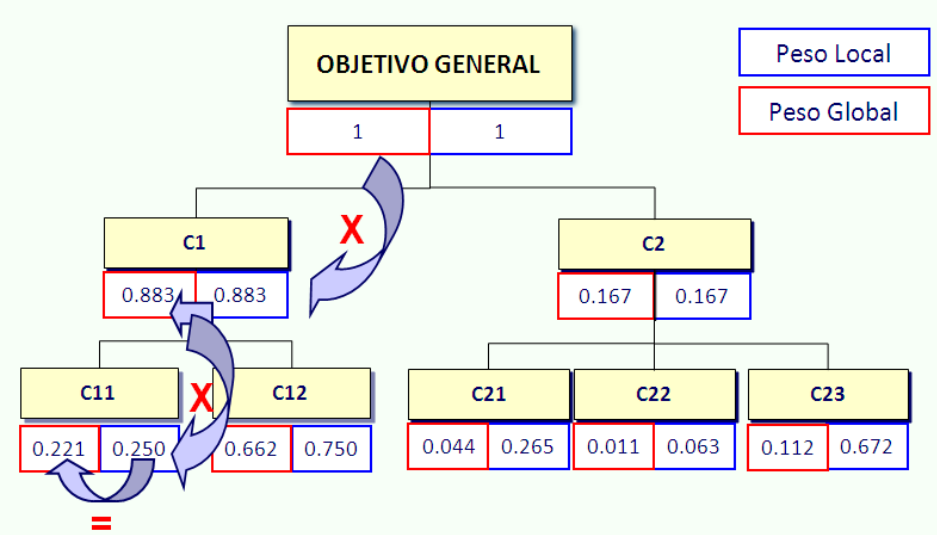


Figura 4.- Cálculo de la puntuación ponderada

	C11	C12	C21	C22	C23
pesos	0,221	0,662	0,044	0,011	0,112
A1	0,316	0,105	0,308	0,143	0,072
A2	0,386	0,258	0,231	0,429	0,649
A3	0,298	0,637	0,462	0,429	0,279

U(A1) = 0,221 x 0,316 + 0,662 x 0,105 + 0,044 x 0,308 + 0,011 x 0,143 + 0,112 x 0,072 = 0,163

U(A2) = 0,221 x 0,386 + 0,662 x 0,258 + 0,044 x 0,231 + 0,011 x 0,429 + 0,112 x 0,649 = 0,344

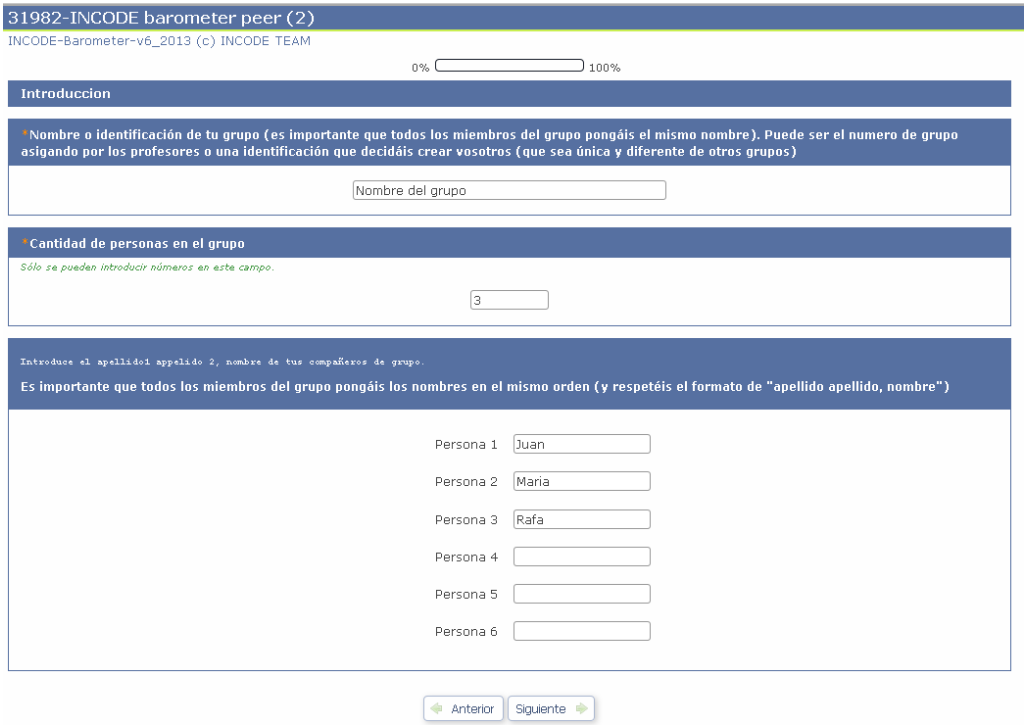
U(A3) = 0,221 x 0,298 + 0,662 x 0,637 + 0,044 x 0,308 + 0,011 x 0,429 + 0,112 x 0,279 = 0,544

Datos

Los datos que analizaremos en este trabajo se corresponden a una asignatura de máster de 5 ECTS, con 10 alumnos organizados en dos grupos de 5 alumnos cada uno, con un sólo profesor impartiendo la asignatura. Los grupos los crearon los propios alumnos para realizar un proyecto de mejora de procesos durante 20 horas presenciales y horas adicionales no presenciales. Analizaremos las notas correspondientes a la competencia de innovación de los estudiantes, articulada como 25 habilidades (crit01 a crit25) agrupadas en 3 categorías de capacidades de innovación (Marin-Garcia et al., 2013), más una pregunta unidimensional sobre la competencia de innovación en general (crit00). Los datos fueron tomados durante el curso 2013-14. Cada alumno recibe las evaluaciones de todos los miembros de su grupo, la puntuación derivada de la media geométrica de las puntuaciones de sus compañeros de grupo (juicio agregado) y la del profesor. En total se han construido 364 matrices (26 matrices por cada alumno, 26matrices de juicios agregados para cada grupo y 26 matrices con la puntuación del profesor para cada grupo).

Los datos se recogieron mediante un cuestionario Web. En la primera pantalla se identificaba el nombre del grupo, el número de componentes y el nombre de los mismos (Figura 5). Era obligatorio que el orden en que se introducían los componentes del grupo fuese el mismo para todos los alumnos del grupo. Se verificó manualmente esta condición al descargar los datos (ningún incumplimiento). A continuación se presentaban varias pantallas, una por cada criterio evaluado, donde los alumnos comparaban las parejas del grupo, incluyéndose ellos mismo respecto a ese criterio (Figura 2). Se ha recogido también el tiempo que tardan los estudiantes en valorar a sus compañeros y el que tarda el profesor en evaluar a los estudiantes.

Figura 5.- pantalla inicial del cuestionario web.



Análisis de la fiabilidad

Para los análisis de fiabilidad *intra-rater* usaremos los Coeficientes de Correlación Intraclase (ICC(k), ICC(C,k), ICC(A,k)) (Morley, 2009; Viladrich Segués y Doval Dieguez, 2011). Se diferencian en el uso de un factor con efectos aleatorios y medida de acuerdo (ICC(k)), dos factores de efectos aleatorios con medida de consistencia (ICC(C,k)) o dos factores de efectos aleatorios y medida de acuerdo (ICC(A,k). En los tres casos, cada objeto/persona recibe el mismo número de evaluaciones. Cada uno de estos coeficientes se puede calcular para estimar la fiabilidad de un evaluador (ICC(1), ICC(C,1), ICC(A,1)) o la fiabilidad de la puntuación calculada como el promedio de las puntuaciones de todos los evaluadores (ICC(k), ICC(C,k), ICC(A,k)).

Calcularemos los ICC usando la macro para SPSS (icc1.sps) proporcionada por Morley (2009), que presenta la ventaja de proporcionar el cálculo de los seis coeficientes ICC simultáneamente. La macro informa de los resultados para una evaluación y para el promedio de evaluaciones. También lo hace, simultáneamente, para múltiples matrices agregadas de acuerdo a los factores elegidos en cada análisis. Además, permite realizar los cálculos aunque haya unos pocos valores perdidos. Los datos, al contrario de los procedimientos estándar de SPSS, deben organizarse de modo que los evaluadores ocupen las filas y las variables evaluadas, así como los factores para agrupar análisis, deben colocarse en columnas. Las variables evaluadas deben nombrarse como QUEST1, QUEST2, etc. Los factores de agrupación pueden tener cualquier nombre siempre que no coincidan con ninguna variable de la macro.

Los valores ICC(C,1) e ICC(C,k) indican la consistencia relativa en las evaluaciones proporcionadas por múltiples evaluadores respecto a diferentes alternativas. Se puede usar para analizar si los evaluadores ordenan las alternativas en un orden consistente con otros evaluadores. No importa tanto si las puntuaciones son iguales, como que el orden relativo de las alternativas sea similar. Los estadísticos ICC(k) e ICC(A,k), representan el grado de acuerdo absoluto o la intercambiabilidad de las puntuaciones de los evaluadores. Es decir, si los evaluadores valoran, con la misma puntuación, las alternativas. Para interpretar estos cuatro ICC se recomienda usar estos niveles de corte: 0,00-0,30 carencia de acuerdo entre evaluadores; 0,31-0,50 acuerdo débil; 0,51-0,70 acuerdo moderado; 0,71-0,90 acuerdo elevado; 0,91-1,00 acuerdo muy elevado (LeBreton y Senter, 2008).

Sin embargo, los valores de ICC(1) e ICC(A,1), que también son estimadores del grado de acuerdo de los evaluadores, se deberían interpretar adoptando los niveles de corte tradicionales para los tamaños de efecto (*effect size*). Así, un valor de 1% debe considerarse como una consistencia pequeña, un valor de 10% puede considerarse una consistencia media y un valor de 25% o superior se puede interpretar como una consistencia elevada (LeBreton y Senter, 2008).

Resultados y discusión

El objetivo principal de la investigación es analizar la consistencia de las puntuaciones en cada dimensión. Por ello, hemos simplificado el problema de ponderación de cada uno de los criterios para que todos los criterios tengan el mismo peso, de modo que la nota ponderada sea el promedio de las notas de cada criterio. Para ellos, se ha establecido que la capacidad de innovación individual tendrá un peso del 48% para el cálculo de la competencia de innovación, la capacidad grupal un 32% y la capacidad de red un 22% (pesos que coinciden con la proporción de criterios dentro de cada categoría). En investigación futura se propone ponderar los criterios usando la comparación pareada.

En la Figura 8 (anexos) se muestra un ejemplo de cálculo de las valoraciones de los alumnos del grupo 2 para el criterio global crit00. Los elementos de las 5 matrices de evaluación realizadas por los alumnos están en las celdas J4:V8. En las celdas J11:V11 están las medias geométricas de las 5 puntuaciones de los alumnos (que representan el juicio agregado del grupo2). Con estos valores se realizan los cálculos de la puntuación de los alumnos normalizadas (J15:N15) y el ratio de consistencia de esta matriz (1,4%). En las celdas D12:H12 se muestran las puntuaciones ponderadas teniendo en cuenta los otros 25 criterios. Así, por ejemplo el Al3 sería la persona de mejor rendimiento en el crit00, según el juicio agregado de todos los componentes del grupo. Del mismo modo también sería el mejor si se consideraran los 25 criterios ponderados. Sin embargo, desde el punto de vista del profesor (D29:H29), el mejor alumno del grupo, tanto en el crit00 como en la puntuación ponderada, es el Al2, siendo el Al3 el segundo (muy cerca del primero y compartiendo el puesto con Al4 y Al5).

Consistencia intra-rater

En la Tabla 6 (anexos) presentamos el Ratio de Consistencia de las 364 matrices construidas. En rojo hemos marcado las celdas cuyos valores de inconsistencia son superiores al nivel recomendado (10%). Una primera lectura nos permite comprobar que las valoraciones del profesor son más consistentes que las de los alumnos. Sin embargo, hay que resaltar que 5 de los 10 alumnos presentan una consistencia excelente en todas las matrices evaluadas (que no sea tan buena como la del profesor, no implica que no pueda ser perfecta para la evaluación de sus compañeros). Además, en determinados criterios, como el Crti04 (evaluar las ventajas y desventajas de las acciones), las valoraciones de los alumnos son mucho más consistentes que las del profesor. También se puede comprobar que, en el agregado de juicios de los alumnos de un grupo, la consistencia es mejor que en la valoración individual de uno solo de sus alumnos y que sus valores son muy próximos a los del profesor (y en cualquier caso excelentes). Esto es importante, porque a pesar de que hay alumnos muy inconsistentes y que, en uno de los grupos, cuatro de los 5 componentes, han tenido problemas de consistencia en varias matrices, en el agregado de juicios se han resuelto las inconsistencias.

En algunas de las matrices del profesor, la inconsistencia es cero porque se trata de un criterio (una matriz) donde los alumnos no han discriminado. Por ejemplo "actitud ética" o "trabajar en grupos multidisciplinares" o criterios similares donde todos los alumnos han sido vistos iguales (en algunos casos porque parece que realmente son iguales y, en otros, porque en las actividades de clase observadas no han dado juego para poner en marcha esa "capacidad" o todos la han puesto en juego de manera muy parecida, no apreciando el profesor diferencias entre ellos). Es normal que los profesores no vean diferencias en algunos criterios porque, en este caso, se observaba a los alumnos un conjunto de 15-20 horas en total a lo largo del curso (actividades de 20' hasta 150 minutos, realizadas durante las clases). Eso significa que a cada grupo se les observa entre 7-10 horas y en cada grupo había 5 alumnos (a los que se observaba a la vez). Sin embargo los alumnos han trabajado juntos, como mínimo, 30 horas durante la asignatura (evidentemente ellos se acaban conociendo mejor de lo que les conoce el profesor). Además, compartían otras 4 asignaturas más en el máster y, aunque en el formulario para evaluarse, se les remarcaba que valoraran lo que hacían en la asignatura solamente, es posible que hayan valorado a la persona en conjunto de todas las asignaturas.

Por otra parte, se puede detectar que hay dos alumnos (Gr01Al03 y Gr02Al10) con muchos problemas de inconsistencia. Estos problemas pueden ser debidos a que se ha tomado la actividad de evaluación con poco rigor, que no hayan comprendido los criterios, que no hayan estado atentos durante la activad de

grupo y, por lo tanto, no sepan como valorar a sus compañeros, a que sean malos evaluadores en general o a algún otro motivo. En cualquier caso, este método nos permite detectar los alumnos con puntuaciones problemáticas y decidir qué hacer (eliminar esos registros del análisis, formar al alumno como evaluador o incluso poner una nota que discrimine la capacidad de evaluación si ese fuese un objetivo de aprendizaje de la asignatura).

Además, podemos comprobar que la mayoría de los criterios no han representado un problema para los diferentes evaluadores, aunque hay algunos criterios que han sido problemáticos para alumnos que, en general, no presentan problemas de consistencia. Por ejemplo crit01 (presenta ideas adecuadas a la tarea) y crit03 (presenta nuevas formas de implantar las ideas). En otras palabras, es posible analizar qué criterios pueden haber resultado más problemáticos en general y cuáles lo son para determinados evaluadores (alumnos o profesor), de modo que se puede incidir en la explicación o proceso de evaluación de esos criterios para mejorar la consistencia de las evaluaciones futuras de los evaluadores.

Consistencia inter-rater

La consistencia entre evaluadores la hemos analizado mediante Coeficientes de Correlación Intraclase aplicados a las puntuaciones obtenidas por cada alumno. Hemos analizado los datos con dos enfoques. En el primero (Figura 6 y Tabla 3) cada fila de la tabla de datos contiene la puntuación obtenida por un alumno, para cada uno de los criterios de evaluación (columnas), para un evaluador concreto. El análisis se hace agrupando los datos de un alumno para todos los evaluadores y comprueba si los evaluadores perciben de manera similar al alumno, usando los 26 criterios. Esto nos permite ver si hay algún alumno difícil de evaluar. Es decir, si hay algún alumno donde los evaluadores dan puntuaciones diferentes usando los 26 criterios como base para la comparación. En general, los índices de consistencia y acuerdo presentan unos valores entre moderado y altos. Los alumnos donde los evaluadores están más de acuerdo son Gr02Al08 y Gr02Al09, mientras que Gr01Al03 es un alumno que los evaluadores muestran escasísimo acuerdo y escasa consistencia.

Figura 6. Tabla parcial de datos A para evaluar consistencia inter-rater según modelo de Morley(2009). Questxx es cada uno de los criterios de evaluación.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	QUEST0	QUEST1	QUEST2	QUEST3	QUEST4	QUEST5	QUEST6	QUEST7	QUEST8	QUEST9	QUEST1	QUEST1	QUEST1	QUEST1	QUEST1	QUEST1	QUEST1	QUEST1	QUEST1	QUEST1	QUEST2	QUEST2	QUEST2	QUEST2	QUEST2	QUEST2	QUEST2	Rater	Rated
2	0,16	0,17	0,26	0,13	0,21	0,15	0,14	0,25	0,26	0,23	0,19	0,23	0,19	0,23	0,19	0,23	0,21	0,16	0,16	0,19	0,22	0,22	0,14	0,23	0,21	0,2	0,2	Agregad	AI01
3	0,1	0,13	0,19	0,19	0,37	0,17	0,15	0,19	0,19	0,19	0,11	0,22	0,13	0,19	0,19	0,19	0,19	0,15	0,15	0,19	0,19	0,19	0,11	0,19	0,19		0,18	Prof	AI01
4	0,09	0,1	0,18	0,05	0,23	0,12	0,15	0,29	0,44	0,27	0,08	0,25	0,21	0,18	0,18	0,23	0,19	0,12	0,12	0,19	0,2	0,12	0,06	0,15	0,18	0,18	0,18	AI01	AI01
5	0,13	0,14	0,16	0,19	0,19	0,19	0,12	0,19	0,19	0,19	0,21	0,19	0,14	0,16	0,19	0,19	0,19	0,19	0,14	0,19	0,19	0,16	0,11	0,21	0,19	0,19	0,18	AI02	AI01
6	0,27	0,31	0,47	0,27	0,17	0,14	0,16	0,24	0,31	0,25	0,42	0,32	0,2	0,39	0,13	0,58	0,23	0,13	0,17	0,16	0,28	0,38	0,28	0,43	0,2	0,26	0,27	AI03	AI01
7	0,15	0,12	0,22	0,09	0,13	0,12	0,1	0,24	0,16	0,22	0,19	0,19	0,19	0,24	0,19	0,11	0,16	0,11	0,16	0,14	0,22	0,16	0,16	0,19	0,19	0,12	0,17	AI04	AI01
8	0,12	0,12	0,2	0,09	0,21	0,12	0,09	0,16	0,16	0,13	0,09	0,13	0,11	0,16	0,16	0,11	0,16	0,16	0,1	0,16	0,12	0,2	0,09	0,11	0,16	0,16	0,14	AI05	AI01
9	0,21	0,25	0,24	0,24	0,22	0,22	0,25	0,22	0,21	0,22	0,27	0,24	0,27	0,16	0,19	0,25	0,25	0,18	0,33	0,19	0,17	0,21	0,19	0,2	0,16	0,21	0,22	Agregad	AI02
10	0,12	0,41	0,27	0,24	0,18	0,12	0,45	0,2	0,18	0,32	0,25	0,25	0,29	0,18	0,18	0,23	0,31	0,12	0,47	0,16	0,11	0,12	0,32	0,18	0,12	0,18	0,23	Gr01AI01	AI02
11	0,25	0,25	0,21	0,19	0,19	0,19	0,25	0,19	0,19	0,19	0,21	0,19	0,28	0,16	0,19	0,19	0,19	0,19	0,28	0,19	0,19	0,21	0,21	0,21	0,19	0,19	0,2	Gr01AI02	AI02
12	0,36	0,21	0,25	0,37	0,31	0,18	0,16	0,19	0,27	0,16	0,19	0,16	0,17	0,14	0,23	0,23	0,3	0,25	0,18	0,16	0,14	0,27	0,09	0,21	0,15	0,16	0,2	Gr01AI03	AI02
13	0,06	0,12	0,11	0,09	0,13	0,25	0,1	0,24	0,16	0,11	0,19	0,19	0,19	0,07	0,09	0,19	0,16	0,11	0,31	0,14	0,11	0,16	0,09	0,1	0,09	0,25	0,15	Gr01AI04	AI02
14	0,34	0,2	0,28	0,31	0,21	0,25	0,28	0,16	0,16	0,26	0,34	0,26	0,26	0,16	0,16	0,29	0,16	0,16	0,26	0,16	0,24	0,2	0,25	0,2	0,16	0,16	0,22	Gr01AI05	AI02
15	0,24	0,25	0,19	0,19	0,18	0,17	0,15	0,19	0,19	0,19	0,25	0,22	0,25	0,19	0,19	0,19	0,19	0,22	0,46	0,19	0,19	0,19	0,22	0,19	0,19		0,21	Prof	AI02

Tabla 3 Resultados ICC basado en alumnos. Análisis según Morley (2009)

Rated	ICC(1)	ICC(C,1)	ICC(A,1)	Promedio-ICC(k)	Promedio-ICC(C,k)	Promedio-ICC(A,k)	MSERROR
Al01	,1006	,2022	,1394	,4474	,6473	,5399	,0031
Al02	,1718	,2203	,1886	,6004	,6717	,6274	,0031
Al03	-,0261	-,0290	-,0298	-,2256	-,2565	-,2653	,0079
Al04	,0973	,1307	,1138	,4357	,5186	,4793	,0057
Al05	,1515	,2322	,1792	,5571	,6806	,6061	,0048
Al06	,0616	,1186	,0916	,3197	,4908	,4195	,0097
Al07	,1735	,2058	,1854	,5993	,6487	,6185	,0118
Al08	,2541	,3253	,2719	,7036	,7706	,7224	,0020
Al09	,3115	,3521	,3207	,7582	,7902	,7659	,0132
Al10	,1064	,2136	,1469	,4522	,6531	,5442	,0020

En el segundo análisis (Figura 7 y Tabla 4), cada fila de la tabla de datos contiene la puntuación en uno de los criterios de evaluación, con los datos de un evaluador, para todos los alumnos evaluados (columnas). El análisis se hace agrupando primero los datos por grupo de alumnos y después por criterio para todos los evaluadores y comprueba si los evaluadores valoran de manera similar el criterio, usando a los 5 alumnos del grupo como sujetos evaluados. Esto nos permite ver donde los evaluadores presentan más acuerdo o consistencia y si hay algún criterio difícil de evaluar. Es decir, si hay algún criterio donde los evaluadores dan puntuaciones diferentes entre sí (Tabla 5).

Figura 7. Tabla parcial de datos B para evaluar consistencia inter-rater según modelo de Morley(2009). Ques-txx es cada uno de los alumnos evaluados.

QUEST1	QUEST2	QUEST3	QUEST4	QUEST5	Rater	Grupo	Sex	Lang	Item
0,16	0,21	0,27	0,16	0,18	Agregad	1			crit00
0,09	0,12	0,18	0,22	0,38	Gr01Al01	1	0	1	crit00
0,13	0,25	0,27	0,11	0,25	Gr01Al02	1	1	0	crit00
0,27	0,36	0,13	0,07	0,17	Gr01Al03	1	0	0	crit00
0,15	0,06	0,34	0,34	0,11	Gr01Al04	1	1	0	crit00
0,12	0,34	0,36	0,11	0,06	Gr01Al05	1	1	1	crit00
0,1	0,24	0,22	0,22	0,22	Prof	1	0	0	crit00
0,17	0,25	0,29	0,19	0,13	Agregad	1			crit01
0,1	0,41	0,41	0,1	0,1	Gr01Al01	1	0	1	crit01
0,14	0,25	0,28	0,16	0,14	Gr01Al02	1	1	0	crit01
0,31	0,21	0,14	0,1	0,17	Gr01Al03	1	0	0	crit01
0,12	0,12	0,25	0,25	0,12	Gr01Al04	1	1	0	crit01
0,12	0,2	0,25	0,28	0,06	Gr01Al05	1	1	1	crit01
0,13	0,25	0,25	0,25	0,13	Prof	1	0	0	crit01
0,26	0,24	0,22	0,16	0,13	Agregad	1			crit02
0,18	0,27	0,27	0,1	0,13	Gr01Al01	1	0	1	crit02
0,16	0,21	0,25	0,14	0,19	Gr01Al02	1	1	0	crit02
0,47	0,25	0,09	0,15	0,1	Gr01Al03	1	0	0	crit02
0,22	0,11	0,22	0,22	0,11	Gr01Al04	1	1	0	crit02
0,2	0,28	0,23	0,12	0,07	Gr01Al05	1	1	1	crit02
0,19	0,19	0,19	0,19	0,19	Prof	1	0	0	crit02

Tabla 4 Resultados ICC basado en criterios. Análisis según Morley (2009)

Grupo	Criterio	ICC(1)	ICC(C,1)	ICC(A,1)	Promedio-ICC(k)	Promedio-ICC(C,k)	Promedio-ICC(A,k)	MSERROR
	crit00	,0640	,0255	,0317	,3235	,1548	,1862	,0640
	crit01	,3566	,3100	,3460	,7951	,7587	,7873	,3566
	crit02	,2586	,2092	,2420	,7094	,6494	,6909	,2586
	crit03	,1375	,0905	,1099	,5199	,4032	,4562	,1375
	crit04	,1037	,0658	,0787	,4476	,3302	,3741	,1037
	crit05	-,0072	-,0314	-,0379	-,0510	-,2612	-,3303	-,0072
	crit06	,3202	,2629	,3048	,7673	,7140	,7542	,3202
	crit07	,1446	,1205	,1324	,5419	,4896	,5166	,1446
	crit08	,1913	,1740	,1844	,6235	,5958	,6128	,1913
	crit09	,1001	,0699	,0806	,4378	,3448	,3801	,1001
	crit10	-,0725	-,0864	-,1044	-,8987	-1,2568	-1,9579	-,0725
	crit11	,3801	,3380	,3709	,8064	,7763	,8002	,3801
	crit12	,4325	,3797	,4232	,8421	,8108	,8370	,4325
1	crit13	,2664	,2242	,2525	,7116	,6625	,6965	,2664
	crit14	,1979	,1797	,1908	,6333	,6053	,6227	,1979
	crit15	,1050	,0712	,0833	,4508	,3492	,3887	,1050
	crit16	,2602	,2464	,2560	,7111	,6959	,7067	,2602
	crit17	,1457	,1168	,1309	,5442	,4807	,5133	,1457
	crit18	,4643	,4090	,4556	,8585	,8289	,8542	,4643
	crit19	,2500	,2275	,2428	,7000	,6733	,6918	,2500
	crit20	,0819	,0459	,0552	,3845	,2518	,2903	,0819
	crit21	-,0934	-,1045	-,1268	-1,4869	-1,9593	-3,7137	-,0934
	crit22	,0139	-,0097	-,0115	,0895	-,0722	-,0862	,0139
	crit23	,1303	,0871	,1049	,5119	,4006	,4507	,1303
	crit24	,1286	,1170	,1226	,5082	,4811	,4945	,1286
	crit25	-,1024	-,1123	-,1278	-1,2582	-1,5354	-2,1243	-,1024
	PONDERADO	,1056	,1356	,1204	,4295	,5000	,4659	,1056
Grupo	Criterio	ICC(1)	ICC(C,1)	ICC(A,1)	Promedio-ICC(k)	Promedio-ICC(C,k)	Promedio-ICC(A,k)	MSERROR
	crit00	,5249	,4694	,5164	,8790	,8533	,8753	,5249
	crit01	,2983	,2476	,2824	,7309	,6777	,7155	,2983
	crit02	,3368	,3269	,3346	,7805	,7727	,7788	,3368
	crit03	,1778	,1353	,1580	,6022	,5228	,5678	,1778
	crit04	,5615	,5298	,5582	,8997	,8875	,8984	,5615
	crit05	,2103	,1712	,1948	,6509	,5912	,6287	,2103
	crit06	,7626	,7243	,7609	,9574	,9484	,9570	,7626
	crit07	,1585	,1365	,1482	,5687	,5254	,5492	,1585
	crit08	,0193	-,0052	-,0061	,1211	-,0374	-,0444	,0193
	crit09	,2421	,2183	,2339	,6846	,6549	,6747	,2421
	crit10	,4329	,4105	,4292	,8424	,8298	,8404	,4329
	crit11	,6982	,6541	,6951	,9402	,9278	,9394	,6982
	crit12	,4485	,4332	,4458	,8468	,8385	,8453	,4485
2	crit13	,2345	,1985	,2217	,6820	,6341	,6660	,2345
	crit14	,2652	,2150	,2487	,7164	,6572	,6986	,2652
	crit15	,3962	,3466	,3863	,8212	,7878	,8151	,3962
	crit16	,2425	,2220	,2358	,6915	,6663	,6835	,2425
	crit17	,0222	-,0054	-,0065	,1371	-,0389	-,0472	,0222
	crit18	,3380	,3015	,3293	,7814	,7513	,7746	,3380
	crit19	-,0511	-,0672	-,0802	-,5162	-,7882	-1,0826	-,0511
	crit20	,0978	,0589	,0711	,4313	,3045	,3487	,0978
	crit21	,2346	,1999	,2223	,6820	,6363	,6668	,2346
	crit22	,4404	,3934	,4324	,8463	,8195	,8421	,4404
	crit23	,3818	,3410	,3730	,8075	,7785	,8016	,3818
	crit24	,5622	,5811	,5639	,8883	,8957	,8890	,5622
	crit25	,3963	,4119	,3999	,7665	,7779	,7691	,3963
	PONDERADO	,5145	,7190	,5365	,8402	,9270	,8517	,5145

En la Tabla 5 resumimos los niveles de consistencia y acuerdo obtenidos en las puntuaciones de los diferentes criterios. En ella podemos observar que la consistencia en el grupo 2 es mayor (sólo las puntuaciones de 4 criterios eran claramente inconsistentes entre evaluadores) que en el grupo 1 (donde hay nueve criterios con nula consistencia entre evaluadores). Salvo el criterio20, el resto de criterios con problemas de consistencia no se repiten. Esto descarta, en principio, que la inconsistencia/desacuerdo entre evaluadores se deba a la formulación del criterio y puede tener su origen en otras causas que habrá que analizar replicando la investigación con otras muestras. En cualquier caso la puntuación ponderada, que es la que se usa en la asignatura para establecer la nota de los alumnos, tiene una consistencia/acuerdo entre evaluadores aceptable o excelente dependiendo del grupo.

Tabla 5.- criterios clasificados en función del acuerdo o fiabilidad de los evaluadores basados en ICC

	Alta	Media	Nula
Grupo 1	Crit01-02	Crit03-04	Crit00
	Crit06	Crit07-08	Crit05
	Crit11-13	Crit14	Crit09-10
	Crit16	Crit17	Crit15
	Crit18-19	Crit23-24	Crit20-22
		Ponderado	Crit25
Grupo 2	Crit00-02	Crit03	Crit08
	Crit04	Crit05	Crit17
	Crit06	Crit07	Crit19-20
	Crit10-12	Crit09	
	Crit14-16	Crit13	
	Crit18	Crit21	
	Crit22-25		
	Ponderado		

Tiempo empleado en evaluar a los compañeros

Por último, hemos recogido, a través de la plataforma web, el tiempo invertido por los evaluadores (alumnos y profesor) para completar las matrices de datos. Partiendo de las 12 evaluaciones realizadas (10 de alumnos y dos del profesor) a los 5 componentes de cada grupo, el tiempo medio invertido fue de 37 minutos (desviación estándar de 21 minutos. Mínimo 19 minutos, máximo 92 minutos). Consideramos que estas cifras son alentadoras pues permiten pensar que la iniciativa es sostenible y no sobrecarga en exceso ni al profesor, ni a los alumnos (que, además, ejercitan su capacidad para evaluar el trabajo de otras personas).

Conclusiones

Hemos presentado un modelo para evaluar las competencias transversales de creatividad, innovación y trabajo en equipo con estudiantes universitarios en asignaturas de máster con pocos alumnos. El modelo ha demostrado ser válido, con una carga de trabajo razonable y asumible por alumnos y profesor, sin excesivas complejidades a la hora de implantarlo y proporcionando unos resultados fiables.

Consideramos que las principales contribuciones de nuestro trabajo son:

- Para los investigadores en el área de AHP, nuestra investigación compara la fiabilidad de las comparaciones pareadas, así como su viabilidad en un contexto poco trabajado (la docencia universitaria).
- Para los investigadores en metodología de evaluación, proporcionamos evidencias de la fiabilidad y validez de un modelo evaluativo basado en decisiones multicriterio.
- Para los docentes universitarios, analizamos las ventajas e inconvenientes de la aplicación del método de decisiones multicriterio AHP para resolver el problema de la evaluación de competencias o habilidades de los alumnos que tienen que ver con el proceso de grupo y no con los resultados o productos del grupo.

Limitaciones e investigación futura:

La principal limitación es que hemos analizado sólo dos grupos y 10 alumnos de máster, lo que nos impide generalizar estos resultados a grupos numerosos o de alumnos de grado. Por otra parte, los alumnos están distribuidos en dos grupos con 5 componentes. Las evaluaciones pareadas de alternativas se han dentro de cada grupo, de modo que las puntuaciones y el orden derivado de las mismas son comparables sólo dentro de un mismo grupo. Es decir, en cada grupo podremos ordenar a las personas de más a menos competente pero no podremos comparar si una persona de un grupo es más o menos competente que una persona de otro grupo. Esta es una de las limitaciones de este trabajo y en investigación futura estableceremos métodos para superar esta limitación.

También hemos considerado el mismo peso para los 25 criterios, de modo que las dimensiones con más criterios, pesan más en la nota ponderada. Por último, en los evaluadores hemos incluido simultáneamente a cada alumno y la agregación de juicios de esos alumnos, que es una medida no independiente de la anterior.

Además, el uso de comparación pareada plantea la dificultada adicional, de integrar en un sólo orden los grupos "disjuntos" en los que dividimos las clases para evitar tener comparaciones pareadas de más de 6 elementos, que se ha demostrado que conducen a evaluaciones inconsistentes en parte por la saturación del evaluador. Sin embargo, esto no es fácilmente adaptable a grupos numerosos (que pueden llegar hasta 90 alumno). Nuestra necesidad como profesores es poder ordenar a todos los alumnos de la clase entre sí, y no solo en el grupo al que pertenecen. Una posibilidad es crear subgrupos con uno o dos alumnos redundantes (por ejemplo en grupos de 10, dos grupos de 6 donde un alumno repite en ambos para ser evaluado por sus compañeros). Otra posibilidad es introducir en las evaluaciones las comparaciones con un "alumno patrón" que permita re-escalar las puntuaciones entre los diferentes grupos. En cualquier caso, tratándose de grupos numerosos, los profesores no estarán en disposición de poder evaluar o comparar parejas de alumnos respecto a las habilidades o comportamientos de las competencias transversales. Sin embargo, si las actividades del curso se han diseñado adecuadamente, los alumnos pueden haber tenido la oportunidad de observar los comportamientos de los compañeros con los que han trabajado y, por lo tanto, están capacitados para evaluarlos. Es, en estos contextos, donde el profesor debería poder apoyarse en las evaluaciones que hacen los alumnos a sus compañeros.

Otras acciones que se pueden llevar a cabo como investigación futura para complementar los hallazgos comentados en este artículo son:

- Comparar la consistencia y acuerdo de la puntuación de un único criterio holístico (crit00) con la ponderación de los 25 criterios o las puntuaciones de una dimensión con los criterios que la constituyen

- Extender a mas grupos, más alumnos y a asignaturas de grado, para poder generalizar las conclusiones
- Analizar la posible reducción de criterios para simplificar el proceso de evaluación. En este sentido se puede empezar por aquellas habilidades muy relacionadas entre si y también por las que han presentado más problemas de consistencia en las evaluaciones.
- Adaptar el modelo y las dimensiones evaluar para ser usadas como herramienta en la evaluación del desempeño en trabajadores de las empresas
- Comparar la consistencia/acuerdo entre las puntuaciones del profesor y la agregación de juicios y las puntuaciones del profesor y las de los alumnos individuales por separado
- Comprobar si los alumnos que son peor evaluadores están asociados con otras variables como rendimiento académico global en la asignatura o la valoración en esta competencia concreta

Agradecimientos

Este trabajo ha sido realizado con la financiación de la Unión Europea ["FINCODA" proyecto 554493-EPP-1-2014-1-FI-EPPKA2-KA]

Anexo- Figura 8.-Ejemplo de la matriz de evaluación de los alumnos del grupo 2 para el criterio00 y las puntuaciones ponderadas en tres casos (puntuación por juicio agregado de alumnos, por el alumno 3, por el profesor)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1				Alumnos evaluados						Crti00	Crti00	Crti00	Crti00	Crti00	Crti00	Crti00	Crti00	Crti00	Crti00	Crti00	Crti00	Crti00	Crti00	Crti00	Crti01	Crti01	Crti01	Crti01	Crti01
2	Evaluador	Grupo	Cantidad	Alt1	Alt2	Alt3	Alt4	Alt5	Alt6	a12	a13	a14	a15	a16	a23	a24	a25	a26	a34	a35	a36	a45	a46	a56	a12	a13	a14	a15	a16
3																													
4	Al3	2	5	Al1	Al2	Al3	Al4	Al5		1	0,5	0,333	0,25		1	0,5	0,333		1	0,5		0,5			0,25	0,25	1	1	
5	Al1	2	5	Al1	Al2	Al3	Al4	Al5		0,5	0,333	2	0,5		1	2	1		2	1		0,5			0,5	0,5	1	1	
6	Al2	2	5	Al1	Al2	Al3	Al4	Al5		1	1	5	2		4	5	2		2	0,5		0,5			3	3	2	1	
7	Al5	2	5	Al1	Al2	Al3	Al4	Al5		3	0,333	0,333	2		0,25	0,25	0,333		1	3		3			1	0,5	0,5	1	
8	Al4	2	5	Al1	Al2	Al3	Al4	Al5		0,333	0,25	1	3		1	4	4		4	4		3			1	0,33	0,25	3	
9																													
10				Puntauciones ponderadas 25 criterios						Puntuaciones crit00															Puntuaciones crit01				
11	Juicio agregado grupo 2		5	Al1	Al2	Al3	Al4	Al5		0,8704	0,425	1,0209	1,0845		1	1,3797	0,9763		1,7411	1,2457		1,0238			0,82	0,57	0,76	1,25	
12				0,2	0,2212	0,2282	0,1837	0,1766	0	1,1666	0,4832	1,7332	1,55		1,45	2,35	1,5332		2	1,8		1,5			1,15	0,92	0,95	1,4	
13								0	ernativa	Al1	Al2	Al3	Al4	Al5									1	ernativa	Al1	Al2	Al3	Al4	Al5
14									mgeo	0,8365	1,0913	1,3854	0,8397	0,9417										mgeo	0,85	1,28	1,46	0,95	0,66425
15								mgeonorm (Wj; Vector B)		0,1642	0,2142	0,2719	0,1648	0,1848								mgeonorm (Wj; Vector B)			0,17	0,25	0,29	0,19	0,13038
16								Vector C (matriz A x Vector b)		0,835	1,083	1,39	0,826	0,935	0							Vector C (matriz A x Vector b)			0,8	1,3	1,4	0,9	0,6532
17								Vector D (C/B)		5,0851	5,0543	5,1105	5,0135	5,058									Vector D (C/B)		5,04	5,04	5,02	5,03	5,00988
18								Lambdamax		5,0643	CI	0,0161	CR	0,014									Lambdamax		5,03	CI	0,01	CR	0,0057
19																													
20	Al3		5	Al1	Al2	Al3	Al4	Al5		1	0,5	0,333	0,25		1	0,5	0,333		1	0,5		0,5			0,25	0,25	1	1	
21				0,1796	0,2335	0,198	0,1516	0,2061	0	1	0,5	0,333	0,25		1	0,5	0,333		1	0,5		0,5			0,25	0,25	1	1	
22								0	ernativa	Al1	Al2	Al3	Al4	Al5									1	ernativa	Al1	Al2	Al3	Al4	Al5
23									mgeo	0,5295	0,6987	1	1,246	2,1694										mgeo	0,57	2,3	2,3	0,57	0,57435
24								mgeonorm (Wj; Vector B)		0,0938	0,1238	0,1772	0,2208	0,3844								mgeonorm (Wj; Vector B)			0,1	0,41	0,41	0,1	0,10177
25								Vector C (matriz A x Vector b)		0,476	0,633	0,902	1,12	1,927	0							Vector C (matriz A x Vector b)			0,5	2	2	0,5	0,5089
26								Vector D (C/B)		5,0716	5,1147	5,0884	5,0708	5,0141									Vector D (C/B)		5	5	5	5	5
27								Lambdamax		5,0719	CI	0,018	CR	0,016									Lambdamax		5	CI	0	CR	0
28	Profesor		5	Al1	Al2	Al3	Al4	Al5		0,333	0,5	0,5	0,5		1	1	1		1	1		1			0,5	0,5	0,5	1	
29				0,1817	0,2096	0,1952	0,1952	0,1952	0														1						
30								0	ernativa	Al1	Al2	Al3	Al4	Al5									1	ernativa	Al1	Al2	Al3	Al4	Al5
31									mgeo	0,5295	1,246	1,1487	1,1487	1,1487										mgeo	0,66	1,32	1,32	1,32	0,65975
32								mgeonorm (Wj; Vector B)		0,1014	0,2386	0,22	0,22	0,22								mgeonorm (Wj; Vector B)			0,13	0,25	0,25	0,25	0,12635
33								Vector C (matriz A x Vector b)		0,511	1,203	1,101	1,101	1,101	0							Vector C (matriz A x Vector b)			0,6	1,3	1,3	1,3	0,6318
34								Vector D (C/B)		5,0376	5,042	5,0066	5,0066	5,0066									Vector D (C/B)		5	5	5	5	5
35								Lambdamax		5,0199	CI	0,005	CR	0,004									Lambdamax		5	CI	0	CR	0

Anexo- Tabla 6. Ratio de Consistencia de cada una de las matrices generadas a partir de los datos

Evaluador-->	Agregado Juicio Gr1	Agregado Juicio Gr2	Gr01 Al01	Gr01 Al02	Gr01 Al03	Gr01 Al04	Gr01 Al05	Gr02 Al06	Gr02 Al07	Gr02 Al08	Gr02 Al09	Gr02 Al10	Prof Gr1	Prof Gr2
CR_crit00	1,3%	2,3%	1,4%	2,0%	4,5%	3,3%	3,6%	8,6%	2,0%	7,7%	6,7%	21,7%	0,4%	2,2%
CR_crit01	0,5%	3,2%	0,0%	1,2%	6,8%	0,0%	4,7%	3,1%	10,1%	4,4%	11,9%	43,2%	0,0%	0,0%
CR_crit02	1,6%	4,9%	2,7%	1,9%	32,3%	0,0%	5,8%	0,0%	5,7%	10,3%	1,5%	110,5%	0,0%	0,0%
CR_crit03	0,5%	0,7%	3,2%	0,0%	7,7%	0,7%	8,1%	4,3%	15,2%	10,1%	7,1%	22,7%	0,0%	0,0%
CR_crit04	1,6%	1,1%	2,7%	0,0%	19,8%	1,1%	0,2%	2,5%	11,1%	2,8%	4,5%	19,6%	8,2%	4,8%
CR_crit05	0,7%	1,0%	0,4%	1,2%	3,6%	0,0%	7,0%	3,7%	3,9%	17,9%	1,6%	24,9%	0,0%	0,0%
CR_crit06	1,6%	1,3%	4,8%	0,0%	32,7%	0,4%	5,5%	0,3%	8,7%	1,2%	6,0%	9,7%	0,0%	0,4%
CR_crit07	0,9%	0,9%	1,9%	0,0%	14,6%	0,8%	0,0%	0,4%	9,0%	4,0%	8,6%	0,0%	0,0%	4,3%
CR_crit08	0,2%	1,6%	1,8%	0,0%	2,6%	0,0%	0,0%	0,5%	1,4%	0,0%	5,2%	32,0%	0,0%	0,0%
CR_crit09	0,3%	1,9%	0,8%	0,0%	9,6%	0,0%	3,2%	1,5%	9,8%	8,1%	5,4%	27,1%	0,0%	0,0%
CR_crit10	0,5%	2,9%	0,2%	0,0%	4,0%	0,0%	8,9%	1,2%	4,6%	3,1%	10,2%	34,7%	0,5%	0,0%
CR_crit11	1,0%	1,1%	0,2%	0,0%	11,3%	0,0%	3,2%	1,7%	4,3%	3,1%	6,8%	33,2%	0,0%	0,2%
CR_crit12	0,5%	3,5%	3,3%	0,0%	2,2%	0,0%	3,7%	0,3%	5,5%	7,4%	8,6%	24,6%	0,0%	0,0%
CR_crit13	0,3%	0,7%	0,0%	0,0%	1,8%	1,4%	0,0%	1,4%	0,0%	6,3%	2,9%	0,0%	0,0%	0,0%
CR_crit14	0,4%	2,0%	0,2%	0,0%	9,6%	0,0%	0,0%	0,0%	2,4%	1,0%	9,1%	14,2%	0,0%	0,0%
CR_crit15	1,1%	0,7%	0,0%	0,0%	6,6%	2,6%	6,0%	0,8%	9,8%	3,1%	6,6%	29,9%	0,0%	0,0%
CR_crit16	0,5%	2,4%	1,5%	0,0%	10,2%	0,0%	0,0%	0,0%	1,0%	1,8%	5,7%	46,4%	0,0%	0,0%
CR_crit17	0,5%	1,9%	0,2%	0,0%	11,2%	0,0%	0,0%	5,1%	1,9%	1,0%	10,0%	43,7%	1,2%	0,3%
CR_crit18	0,7%	0,6%	0,0%	0,0%	8,4%	0,2%	2,2%	0,0%	6,1%	3,8%	3,6%	0,0%	0,0%	0,0%
CR_crit19	0,4%	1,3%	2,4%	0,0%	4,2%	0,0%	0,0%	0,0%	1,2%	2,4%	1,9%	20,6%	0,0%	0,0%
CR_crit20	0,5%	1,2%	0,3%	0,0%	6,7%	0,0%	1,2%	1,1%	2,5%	1,1%	1,7%	20,6%	0,0%	0,0%
CR_crit21	0,3%	0,4%	0,2%	3,9%	3,6%	0,0%	0,2%	2,4%	7,9%	5,7%	4,8%	0,0%	0,0%	1,2%
CR_crit22	0,1%	2,7%	1,5%	0,0%	8,8%	0,2%	4,3%	2,5%	3,9%	2,5%	1,9%	49,8%	0,0%	0,0%
CR_crit23	0,4%	0,3%	2,7%	0,0%	5,6%	0,3%	5,5%	7,2%	3,1%	1,5%	4,9%	0,0%	0,0%	0,0%
CR_crit24	0,3%	0,7%	1,5%	0,0%	7,9%	0,0%	0,0%	3,0%	3,9%	0,0%	1,1%	0,0%	0,0%	0,0%
CR_crit25	0,1%	0,5%	0,0%	0,0%	3,1%	0,0%	0,0%	0,5%	0,9%	0,0%	6,1%	0,0%	0,0%	0,0%

Referencias

Albayrak, E.; Erensal, Y. C. (2004). Using analytic hierarchy process (AHP) to improve human performance: An application of multiple criteria decision making problem. Journal of Intelligent Manufacturing, Vol. 15, nº. 4, pp. 491-503.

Andreu Andrés, M. A.; García-Casas, M. (2014). La evaluación de la participación en equipos de trabajo universitarios (Assessment of participation in higher education team working activities). WPOM-Working Papers on Operations Management, Vol. 5, nº. 1.

Chin, S.; Pun, F.; Xu, Y.; Chan, S. F. (2002). An AHP based study of critical factors for TQM implementation in Shanghai manufacturing industries. Technovation, Vol. 22, nº. 11, p. 707.

Comisión Europea (1995). Libro verde de la innovación. Comisión Europea (ES/13/95/55220800.P00).

Conlon, P.; Hecker, K.; Sabatini, S. (2012). What should we be selecting for? A systematic approach for determining which personal characteristics to assess for during admissions. BMC Medical Education, Vol. 12, nº. 1.

De Beuckelaer, A.; Toonen, S.; Davidov, E. (2013). On the optimal number of scale points in graded paired comparisons. Quality and Quantity, Vol. 47, nº. 5, pp. 2869-2882.

European Commission (2008). The European Qualifications Framework for Lifelong Learning (EQF). Office for Official Publications of the European Communities.

Fernández March, A. (2010). La evaluación orientada al aprendizaje en un modelo de formación por competencias en la educación universitaria. Revista de Docencia Universitaria, Vol. 8, nº. 1, pp. 11-34.

Gee, S. (1981). Technology transfer, innovation & international cometitiveness. Wiley & Sons.

Goffin, K.; Mitchell, R. (2010). Innovation management. Palgrave-MacMillan.

González Pernía, J. L.; Peña-Legazkue, I. (2007). Determinantes de la capacidad de innovación de los negocios emprendedores en España. Economia Industrial nº. 363, pp. 129-147.

Goswami, S.; Mathew, M. (2005). Definition of innovation revisited: An empirical study on Indian information technology industry. International Journal of Innovation Management, Vol. 09, nº. 03, pp. 371-383.

Hatzinger, R.; Dittrich, R. (2012). Prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. Journal of Statistical Software, Vol. 48, nº. 10.

Ikehara, K.; Toyoda, H. (2012). A course evaluation model using paired comparisons to integrate importance of rating criteria and students' ratings of teaching: Comparison of students' and teachers' evaluations. Japanese Journal of Educational Psychology, Vol. 60, nº. 1, pp. 48-59.

Ingols, C.; Shapiro, M. (2014). Concrete Steps for Assessing the Soft Skills in an MBA Program. Journal of Management Education, Vol. 38, nº. 3, pp. 412-435.

Kan Ma, H.; Min, C.; Neville, A.; Eva, K. (2013). How Good Is Good? Students and Assessors' Perceptions of Qualitative Markers of Performance. Teaching and Learning in Medicine, Vol. 25, nº. 1, pp. 15-23.

Kasirian, M. N.; Yusuff, R. M.; Ismail, M. Y. (2010). Application of AHP and ANP in supplier selection process-a case in an automotive company. International Journal of Management Science and Engineering Management, Vol. 5, nº. 2, pp. 125-135.

Khalaf, M. A. & El Mokadem, M. Y. (2011). A Systematic approach for prioritizing lean practices using AHP, EurOMA.

Klippel, A. F.; Petter, C. O.; Antunes, J. (2008). Management Innovation, a way for mining companies to survive in a globalized world. Utilities Policy, Vol. 16, nº. 4, pp. 332-333.

Lasnier, F. (2000). Réussir la formation par compétences. Guérin.

Lawson, B.; Samson, D. (2001). Developing innovation capability in organizations: A dynamic capabilities approach. International Journal of Innovation Management, Vol. 05, nº. 03, pp. 377-400.

LeBreton, J. M.; Senter, J. L. (2008). Answers to 20 Questions About Interrater Reliability and Interrater Agreement. Organizational Research Methods, Vol. 11, nº. 4, pp. 815-852.

Lehto, A.; Kairisto-Mertanene, L.; Penttilä, T. (2011). Towards innovation pedagogy. A new approach to teaching and learning for universities of applied sciences. Turku University of Apllied Sciences.

Li, Y., Wang, L., & Dong, X. (2009). An AHP-based study on the assessment of bilingual teaching in higher education institute, in 2009 2nd International Conference on Education Technology and Training, ETT 2009; Sanya; China; 13 December 2009 through 14 December 2009; Category numberP3936; Code 79728, 2009 2nd International Conference on Education Technology and Training, ETT 2009; Sanya; China; 13 December 2009 through 14 December 2009; Category numberP3936; Code 79728, pp. 231-234.

Lind, D. S.; Rekkas, S.; Bui, V.; Lam, T.; Beierle, E.; Copeland, I. I. I. (2002). Competency-Based Student Self-Assessment on a Surgery Rotation. Journal of Surgical Research, Vol. 105, nº. 1, pp. 31-34.

Ljungman, A. G.; Silén, C. (2008). Examination involving students as peer examiners. Assessment & Evaluation in Higher Education, Vol. 33, nº. 3, pp. 289-300.

Lohmann, J. R.; Rollins, H. A.; Hoey, J. J. (2006). Defining, developing and assessing global competence in engineers. European Journal of Engineering Education, Vol. 31, nº. 1, pp. 119-131.
Lyons, R. K.; Chatman, J. A.; Joyce, C. K. (2007). Innovation in services: Corporate culture and investment banking. California Management Review, Vol. 50, nº. 1, pp. 174-191.
Marin-Garcia, J. A. (2009). Los alumnos y los profesores como evaluadores. Aplicación a la calificación de presentaciones orales. Revista Española de Pedagogía, Vol. 67, nº. 242, pp. 79-97.
Marin-Garcia, J. A.; Aznar-Mas, L. E.; González-Ladrón-de-Gevara, F. (2011). Innovation types and talent managment for innovation. Working Papers on Operations Management, Vol. 2, nº. 2, pp. 25-31.
Marin-Garcia, J. A.; Garcia-Sabater, J. P.; Miralles, C.; Rodríguez Villalobos, A. (2008). Profile and competences of Spanish industrial engineers in the European Higher Education Area (EHEA). Journal of Industrial Engineering and Management, Vol. 1, nº. 2, pp. 269-284.
Marin-Garcia, J. A.; Garcia-Sabater, J. P.; Perello-Marin, M. R.; Canos-Daros, L. (2009a). Proposal of skills for the bachelor degree of Industrial Engineering in the context of the new curriculum. Intangible Capital, Vol. 5, nº. 4, pp. 387-406.
Marin-Garcia, J. A.; Garcia-Sabater, J. J.; Bonavia, T. (2009b). The impact of Kaizen Events on improving the performance of automotive components' first-tier suppliers. International Journal of Automotive Technology and Management, Vol. 9, nº. 4, pp. 362-376.
Marin-Garcia, J. A.; Perez-Peñalver, M. J.; Watts, F. (2013). How to assess innovation competence in services: The case of university students. Dirección y Organización nº. 50, pp. 48-62.
Marrin, M. L.; McIntosh, K. A.; Keane, D.; Schmuck, M. L. (2004). Use of the paired-comparison technique to determine the most valued qualities of the McMaster medical programme admissions process. Advances in Health Sciences Education, Vol. 9, nº. 2, pp. 129-135.
Mol, M. J.; Birkinshaw, J. (2009). The sources of management innovation: When firms introduce new management practices. Journal of Business Research, Vol. 62, nº. 12, pp. 1269-1280.
Morley, D. D. (2009). SPSS macros for assessing the reliability and agreement of student evaluations of teaching. Assessment & Evaluation in Higher Education, Vol. 34, nº. 6, pp. 659-671.
Mula, J.; Días-Madroñero, M.; Poler, R. (2012). Configuración del Grado en Ingeniería de Organización Industrial en las universidades españolas. Dirección y Organización, Vol. 47, pp. 5-20.
Penttilä, T. & Kairisto-Mertanene, L. (2012). Innovation competence barometer ICB - a tool for assessing students' innovation competences as learning outcomes in higher education, in INTED2012 Conference. 5th-7th March 2012, pp. 6347-6351.
Perrenoud, P. (2005). La universitat entre la transmissió de coneixements i el desenvolupament de competències. El debat sobre les competències a l'ensenyament universitari , ICE UB. Documents de Docència Universitària, núm. 5., pp. 8-25.
Pond, K. (2007). Student Experiences of Peer Review Marking of Team Projects. International Journal of Management Education, Vol. 6, nº. 1, pp. 30-43.
Saaty, T. L. (1980). The Analytic Hierarchy Process. McGraw-Hill.
Schumpeter, J. (1934). The Theory of Economic Development. Harvard University Press.
Terry, R. E.; Harb, J. N.; Hecker, W. C.; Wilding, W. V. (2002). Definition of student competencies and development of an educational plan to assess student mastery level. International Journal of Engineering Education, Vol. JAM-PDF, nº. 2, pp. 225-235.
Tonnessen, T. (2005). Continuous innovation through company wide employee participation. TQM Magazine, Vol. 17, nº. 2, pp. 195-207.
Vaccaro, I. G.; Jansen, J. J. P.; Van Den Bosch, F. A. J.; Volberda, H. W. (2012). Management Innovation and Leadership: The Moderating Role of Organizational Size. Journal of Management Studies, Vol. 49, nº. 1, pp. 28-51.
Van Overveld, K. & Verhoeff, T. (2013). Self-consistent peer ranking for assessing student work: Dealing with large populations, in CSEDU 2013 - Proceedings of the 5th International Conference on Computer Supported Education, pp. 399-404.
Veronese Bentes, A.; Carneiro, J.; Ferreira da Silva, J.; Kimura, H. (2012). Multidimensional assessment of organizational performance: Integrating BSC and AHP. Journal of Business Research, Vol. 65, nº. 12, pp. 1790-1799.
Viladrich Segués, M. C.; Doval Dieguez, E. (2011). Medición: fiabilidad y validez. Laboratori d'Estadística Aplicada i de Modelització (UAB).
Villa Sánchez, A.; Poblete, M. (2007). Aprendizaje basado en competencias. Una propuesta para la evaluación de las competencias genéricas. Universidad de Deusto.
Watts, F.; Garcia-Carbonell, A.; Andreu Andrés, M. A. (2013). Innovation competencies development: INCODE barometer and use guide. Turku University od Applied Sciences.

Watts, F.; Marin-Garcia, J. A.; Garcia-Carbonell, A.; Aznar-Mas, L. E. (2012). Validation of a rubric to assess innovation competence. Working Papers on Operations Management, Vol. 3, nº. 1, pp. 61-70.

Yao, J. (2010). Feature AHP method used for excellent students evaluation, in 2010 International Conference on Optics, Photonics and Energy Engineering, OPEE 2010; Wuhan; China; 10 May 2010 through 11 May 2010; Category numberCFP1033I-PRT; Code 81252, 2010 International Conference on Optics, Photonics and Energy Engineering, OPEE 2010; Wuhan; China; 10 May 2010 through 11 May 2010; Category numberCFP1033I-PRT; Code 81252, pp. 422-425.

Yiu, E.; Chan, K.; Mok, R. (2007). Reliability and confidence in using a paired comparison paradigm in perceptual voice quality evaluation. Clinical Linguistics and Phonetics, Vol. 21, nº. 2, pp. 129-145.