



El efecto del empleo de anclas verbales o numéricas en una rúbrica para evaluar presentaciones orales, sobre la consistencia de las puntuaciones de los alumnos universitarios

Juan A. Marin-García^a, Cristina Santandreu-Mascarell^b, Julien Maheut^a y Gregorio Rius Sorolla^b

^aROGLE. Dpto. de Organización de Empresas. Universitat Politècnica de València. Camino de Vera S/N 46021 Valencia. jamarin@omp.upv.es; julien.maheut@gmail.com, ^bDpto. de Organización de Empresas. Universitat Politècnica de València. Camino de Vera S/N 46021 Valencia. crisanma@omp.upv.es; greriuso@upv.es

Abstract

In this paper we compare the use of two versions of a rubric for evaluation by university students, of oral presentations of their peers. We have analyzed two sets of data from Industrial Engineering students: 29 presentations (449 independent evaluations) in 2004 and 36 presentations (468 independent evaluations) in 2013. Contrary to expectations the consistency of assessments using improved rubric (second version) are worse than the first. But this may be due to differences in how to perform the evaluation activity with students.

Keywords: oral presentations, rubric, peer-assessment, Higher Education, consistency, reliability, Psychometric analysis

Resumen

En este trabajo comparamos el uso de dos versiones de una rúbrica para la evaluación, por parte de alumnos universitarios, de las presentaciones orales de sus compañeros. Hemos analizado dos conjuntos de datos provenientes de estudiantes de Ingeniería de Organización Industrial: 29 presentaciones (449 evaluaciones independientes) en el año 2004 y 36 presentaciones (468 evaluaciones independientes) en el año 2013. Contrariamente a lo esperado la consistencia de las evaluaciones de la rúbrica mejorada (segunda versión) son peores que la primera. Pero esto puede ser debido a las diferencias en el modo de realizar la actividad de evaluación con los alumnos.

Palabras clave: presentaciones orales, rúbrica, evaluación por pares, universidad, consistencia, fiabilidad, análisis psicométrico.



Introducción

Una de las competencias transversales citadas con más frecuencia en la educación superior es la relacionada con las presentaciones en público (De Grez, Valcke, & Roozen, 2012; García-Ros, 2011; Langan et al., 2005; Marin-Garcia, 2009; Watts, García-Carbonell, & Llorens, 2006).

Para la evaluación de presentaciones orales se pueden usar criterios objetivos o subjetivos. Los criterios objetivos son aspectos observables independientes del observador (por ejemplo, hacia dónde mira la persona que expone, si se cumple el tiempo, si existen gráficos o dibujos en la presentación, etc.). Los criterios subjetivos dependen del observador (por ejemplo si gusta el formato o no, si el tema resulta interesante, etc.)

Existe cierta cantidad de trabajos que proponen rúbricas para evaluar esta competencias (Andreu, 2005; García Ros & et.al., 2012; Magin & Helmore, 2001; Marin-Garcia, 2009), pero pocos trabajos las validan comprobando su fiabilidad y muchos menos reutilizan las rúbricas validadas por otros autores para comprobar cuál es el nivel de desempeño alcanzado por sus estudiantes en esta competencia (Marin-Garcia & Santandreu-Mascarell, 2015; van Ginkel, Gulikers, Biemans, & Mulder, 2015).

Por otra parte, no hemos encontrado ninguna investigación donde se comente el efecto del modo de plantear las evaluaciones (si se introduce algún elemento para que al alumno le interese tomarse en serio las puntuaciones de sus compañeros) y formato de las anclas de los diferentes niveles de la rúbrica.

Objetivos

Partiendo del trabajo de Marin-Garcia (2009) (*Figura 1*), en esta comunicación planteamos:

1. Mejorar la validez de contenido de la rúbrica versión original (con © el año 2005) para criterios objetivos de evaluación de presentaciones orales, creando la versión 2012 (*Figura 3*)
2. Comparar las propiedades psicométricas de las dos versiones de la rúbrica y el posible efecto de las diferencias de aplicación

Guía para evaluación de exposiciones orales - criterios OBJETIVOS ©Juan Antonio Marin García 2005 Dept. Organización de Empresas – ETSII – jamarin@omp.upv.es					Construido colaborativamente con los alumnos de 7102: tormenta de ideas, filtro objetivos/subjectivos, selección (dos grupos: mañana (20 alumnos) y tarde (7 alumnos)). La					
Criterio	Niveles/puntos				Evaluación de la Exposiciones					
	0	1	2	3						
Mirada	Solo lee notas o transparencia	Mira a una sola persona o a unos pocos	mira a los ojos a todos alguna vez	mira a todos a los ojos con frecuencia						
Aparenta tranquilidad	NO	Poco	---	SI						
Al hablar	No se oye	Se oye pero habla demasiado rápido/lento	Se oye pero habla un poco rápido/lento	Se oye y ritmo adecuado						
Acetatos Legibles	No, es bastante difícil leerlos	Cuesta un poco	--	SI						
Participan las dos personas en la exposición	No	Se nota mucha diferencia entre una y otra persona	--	SI						
Subtotal A										
Acetatos	No incluye dibujos ni colores	Tiene más de un color	Incluye gráficos o dibujos	Colores, gráficos y dibujos						
Acetato esquematizado	No, párrafos largos	SI, pero más de 14 líneas de texto	--	SI, 14 líneas de texto o menos por acetato						
Terminan antes de los 3 minutos	No, el profesor les debe avisar	--	--	SI						
Exposición centrada en aspectos de la asignatura	El tema no pertenece a la asignatura o no cumple los requisitos de la actividad	--	--	SI						
Subtotal B										
Puntos totales: Subtotal A+ Subtotal B (máximo 27 puntos)										

Figura 1. Versión rúbrica con anclas verbales, desarrollada en año 2004 con © del año 2005

Desarrollo de la innovación

Hemos creado la versión 2012 (Figura 3) añadiendo nuevos criterios a la versión 2004-2005 (Figura 1) y clarificando los niveles de puntuación, reformulando el contenido del ancla verbal original y añadiendo nuevas anclas verbales para niveles intermedios. Para este trabajo hemos aprovechado el feedback recibido tras varios años de uso de la rúbrica con alumnos de grado y máster. También se han analizado las rúbricas publicadas (Andreu, 2005; Dunbar, Brooks, & Kubicka-Miller, 2006; García-Ros, 2011; Magin & Helmore, 2001; Watts et al., 2006), así como los comentarios de compañeros del grupo EICE-IEMA a través de diferentes reuniones de trabajo para mejorar la rúbrica.

Para la consistencia inter-rater usaremos las puntuaciones de diferentes evaluadores para una misma presentación oral (Marin-Garcia, 2016). Utilizaremos dos conjuntos de datos para comparar las propiedades psicométricas de las rúbricas:

1. Para la construcción de la versión 2004-2005 los criterios a evaluar fueron seleccionados por los estudiantes. La tarea del profesor fue integrar las visiones de los diferentes grupos y generar la versión definitiva de la parrilla, incluyendo unas guías de puntuación para cada criterio. La parrilla contenía 9 criterios y cada uno de ellos podía puntuarse entre 0 y 3. Por lo tanto, la nota máxima para una exposición eran 27 puntos. Los alumnos objeto de estudio estaban matriculados en una de las dos clases del tercer curso de Ingeniero de Organización Industrial. Ninguno de los alumnos había participado en actividades de evaluación de compañeros con anterioridad a esta asignatura. La actividad evaluada se realizaba en parejas. Consistía en realizar una entrevista a dos mandos de empresa, comparar sus respuestas con la teoría impartida en el curso y presentar el resultado

de su trabajo ante los demás compañeros en clase. La actividad se exponía el último día del curso y no era obligatoria, aunque puntuaba un 10% en la nota final de la asignatura - 5% por la nota de la exposición (promedio de la nota de los compañeros y del profesor) y 5% por el grado de acuerdo de las notas puestas por cada alumno al compararla con el promedio de notas de todos los alumnos). Se realizaron 29 presentaciones. El día de la presentación, los alumnos puntuaban a sus compañeros utilizando una copia en papel de la rúbrica (*Figura 1*). Con el fin de no saturar a los alumnos y que pudieran prestar atención también al contenido de las presentaciones, los alumnos evaluaban sólo una de cada cuatro presentaciones. La asignación de qué exposición debían evaluar se hizo en función del lugar que ocupaban en las mesas de clase (bancos con 4 asientos), por lo que podemos considerar que fue una asignación al azar. Los datos provienen de la investigación publicada por Marin-Garcia (2009) pero se analizarán con un nuevo procedimiento estadístico. En total disponemos de 449 puntuaciones diferentes.

2. Para la versión 2012, los datos se obtuvieron de 21 presentaciones en la sesión del grupo de mañana y 15 en la de tarde. Los datos de las 468 evaluaciones independientes se obtuvieron en la sesión del 17/12/2013, donde los alumnos realizaban una presentación oral del lanzamiento del MRP (Material Requirement Planning), a partir de los datos desarrollados por cada grupo. Al inicio de la sesión de las presentaciones, se les informó a los alumnos de que debían evaluar las exposiciones de sus compañeros, mediante la rúbrica que les habíamos facilitado en las sesiones previas. La rúbrica se rellenaba, durante la exposición, en una aplicación web usando los ordenadores del aula y los alumnos podían escribir comentarios o recomendaciones en un campo de texto libre en la hoja de la rúbrica. Se les comentó que dicha evaluación sería anónima y que no formaría parte de la nota de las presentaciones de sus compañeros. Se les instó a que puntuaran con responsabilidad y que los comentarios críticos fuesen constructivos. Recordando, que todos iban a ser evaluados por distintos compañeros y que el resultado de sus evaluaciones, solo las recibiría el evaluado al finalizar todas las presentaciones. En cada grupo exponían las tres personas participantes, cada uno disponía de un máximo de 3 minutos para sus presentaciones (10 minutos en total cada grupo). Después de cada presentación de grupo, se dedicaban 5-10 minutos para permitir a los alumnos finalizar sus evaluaciones si se les había quedado alguna incompleta.

Tabla 1.- Similitudes y diferencias entre las dos aplicaciones

	Conjunto de datos 1	Conjunto de datos 2
Rúbrica usada	Versión 2004-2005	Versión 2012
Diseño de la rúbrica	Los alumnos eligen los criterios y el profesor establece los niveles y sus anclas	El profesor elige los criterios, los niveles y sus anclas
Número de criterios	9	11

Niveles	4 (0-3)	4 (0-3)
Administración	Rubrica (versión con anclas verbales) conocida un mes antes por los alumnos. Uso de pruebas en un simulacro de presentación.	Rubrica (versión con anclas verbales) conocida un mes antes por los alumnos.
Puntuación	Identificación del evaluador para el profesor Usan versión en papel con las anclas verbales. Cada evaluador evalúa una presentación y descansa 3	Evaluador anónimo Usan versión web solo con las anclas numéricas. Todos los evaluadores evalúan todas las presentaciones de la sesión
Control de responsabilidad	Sensibilización de los alumnos sobre la calidad del feedback a sus compañeros. La mitad de la nota de la actividad depende que hayan sido buenos evaluadores	Sensibilización de los alumnos sobre la calidad del feedback a sus compañeros.
Alumnos	3º Ingeniero de organización	3º Ingeniero de organización (mismo centro que datos 1)
Número de presentaciones	29	36
Tiempo máximo de cada presentación	3 minutos	3 minutos
Asignatura	Gestión	Dirección de Operaciones

Para los análisis de fiabilidad usaremos el procedimiento explicado por Marin-Garcia et al. (2014) (EPIDAT, 2014; LeBreton & Senter, 2008; Morley, 2009; Viladrich Segué & Doval Dieguez, 2011).

En la *Figura 2* mostramos la tabla de datos. 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 una presentación para todos los evaluadores y comprueba si los evaluadores perciben de manera similar al alumno, usando los 9 u 11 criterios de la rúbrica (según versión). Esto nos permite ver si hay algún alumno donde los evaluadores dan puntuaciones diferentes usando todos criterios como base para la comparación.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	QUEST1	QUEST2	QUEST3	QUEST4	QUEST5	QUEST6	QUEST7	QUEST8	QUEST9	QUEST10	QUEST11	QUEST12	QUEST13	QUEST14	QUEST15	Rater	Rated
2	2,00	3,00	3,00	2,00	3,00	3,00	3,00	3,00	3,00	2,00	3,00	2,50	3,00	2,67	2,73	profesor	1
3	3,00	2,00	3,00	3,00	3,00	3,00	3,00	3,00	3,00	2,00	3,00	2,75	3,00	2,67	2,82	profesor	2
4	2,00	3,00	3,00	2,00	3,00	3,00	3,00	3,00	3,00	2,00	2,00	2,50	3,00	2,67	2,64	AI02	1
5	2,00	3,00	2,00	2,00	2,00	3,00	3,00	3,00	3,00	2,00	3,00	2,25	2,67	2,67	2,55	AI01	2

Figura 2. Tabla parcial de datos A para evaluar consistencia inter-rater según modelo de Morley(2009). Questxx es cada uno de los criterios de evaluación.

Resultados

En la *Figura 3* mostramos la versión mejorada de la rúbrica, que fue distribuida a los alumnos del segundo conjunto de datos un mes antes de las presentaciones.

Criterio	Niveles/puntos				Evaluación de la Exposición #														
	0	1	2	3															
Toma en cuenta de la audiencia	Ignora a la audiencia o sobreactúa de modo innecesario	Frialdad. No provoca ninguna reacción en la audiencia	Consigue la atención al inicio	Hace participar, cuenta anécdotas, mantiene el "enganche"															
Mirada	Sólo lee notas o el contenido de la transparencia	Mira a una sola persona o a unos pocos	Mira a los ojos a todos alguna vez	Mira a los ojos a todos con frecuencia															
Control de los nervios	Bloqueo total. Afecta a la comunicación	Temblores (voz o mano)	Busca empatía de audiencia (o salidas antes ante nerviosismo)	Aparenta tranquilidad															
Movimiento	Se desplaza y hace gestos que despidan	No se mueve	Domina el espacio y se expresa con las manos	El lenguaje corporal refuerza el mensaje															
Volumen / Velocidad (voz)	No se le oye	Se le oye pero habla demasiado rápido / lento	Se le oye pero habla un poco rápido / lento	Volumen y velocidad adecuada															
Expresión en la voz	Incoherente, despista y aleja del mensaje	Lineal, sin énfasis	Modula y enfatiza	Modula y enfatiza con coherencia con el mensaje															
Transmisión de mensaje	Confuso, poco claro	Cierta estructura y objetivos	Simple y claro	Remarca las ideas principales															
TEXTO Ayudas Visuales legibles	Bastante difícil leerlas	Difíciles de leer.	Información no pertinente o excesiva.	Legibles y congruentes con el mensaje															
TEXTO Ayudas Visuales esquematizadas	Párrafos largos. Demasiada información	Información inconexa	14 líneas de texto o menos por transparencia	Información estructurada															
AYUDAS VISUALES (dibujos y uso del color)	Imágenes inadecuadas o sin sentido	No incluye dibujos ni colores. Sólo hay texto	Uso de color o gráficos para resaltar información.	Colores y gráficos o dibujos coherentes con el objetivo del presentador															
Control del tiempo	Sobrepasa > de 5 minutos el tiempo asignado	Sobrepasa < 5 minutos	---	Termina a tiempo o antes del tiempo asignado															
Puntos totales: (máximo 33 puntos)																			

Rúbrica para evaluación de exposiciones orales - criterios OBJETIVOS - © Juan A. Marín-García 2012 -

 This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License](http://creativecommons.org/licenses/by-nc-nd/2.5/). Es libre de copiar y distribuir este texto según los términos de esta licencia -pero sólo para usos no comerciales-. Cualquier uso de este documento deberá citar al autor. El texto completo de la licencia puede consultarse en la URL <http://creativecommons.org/licenses/by-nc-nd/2.5/>

Figura 3. Versión 2012 con anclas verbales

Durante las exposiciones del año 2013, los alumnos introducían los datos en un formulario web donde sólo aparecía el nombre del criterio y los números de los diferentes niveles:

Encuesta MRP GRUPO 1 Editar este formulario

Valoración de la presentación de la práctica MRP

***Obligatorio**

PONENTE ESTHER *

	0	1	2	3
Toma en cuenta a la audiencia	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mirada	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Control de los nervios	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Movimiento	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Volumen / Velocidad (voz)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Expresión en la voz	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Transmisión de mensaje	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
TEXTO Ayudas Visuales legibles	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
TEXTO Ayudas Visuales esquematizadas	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
AYUDAS VISUALES Dibujos y uso del color)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Control del tiempo	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Comentarios o recomendaciones (ponente Esther):

Figura 4. Formulario web para la versión 2012 con anclas numéricas

En general (Tabla 2, Figura 5 y Figura 6), los índices de consistencia y acuerdo presentan unos valores bastante más elevados en el conjunto de datos 1 que en el 2. Si nos centramos en los ICC que representan la consistencia o acuerdo del promedio de las puntuaciones de varios evaluadores en el diseño de efectos aleatorios (ICC(C,k) e ICC (A,k), para la serie de datos 1, prácticamente en todos los casos (salvo en 3) se supera el valor de 0,7. Además los valores de consistencia y acuerdo son prácticamente iguales. De modo que los alumnos no solo ordenan igual las presentaciones mejores y peores sino que les otorgan unos valores de puntuación prácticamente iguales. Sin embargo, en el conjunto de datos 2, en menos de la mitad de las presentaciones se consiguen valores de 0,7 en el índice de consistencia. Además, los valores de acuerdo son sensiblemente menores en este caso. Todo ello hace pensar que bien la rúbrica 2012 o bien el contexto en el que se utilizó es mucho menos fiable.

Tabla 2.- resultados replicados de la investigación Marin-Garcia (2009)

Etiquetas de fila	Promedio de ICC(1)	Promedio de ICC(C,1)	Promedio de ICC(A,1)	Promedio de ICC(k)	Promedio de ICC(C,k)	Promedio de ICC(A,k)
Datos 1	0,46	0,48	0,46	0,86	0,87	0,86
Datos 2	0,07	0,15	0,10	0,29	0,62	0,51

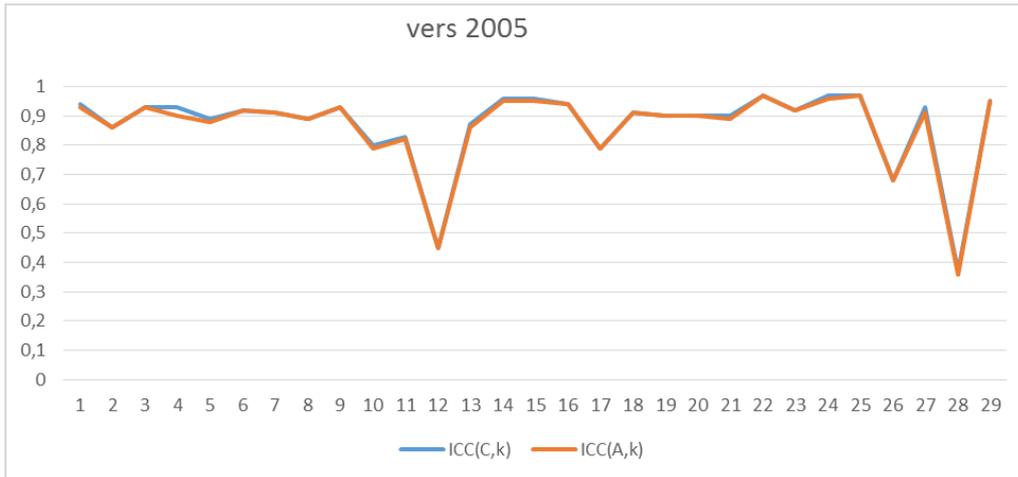


Figura 5. Resultados ICC conjunto de datos 1

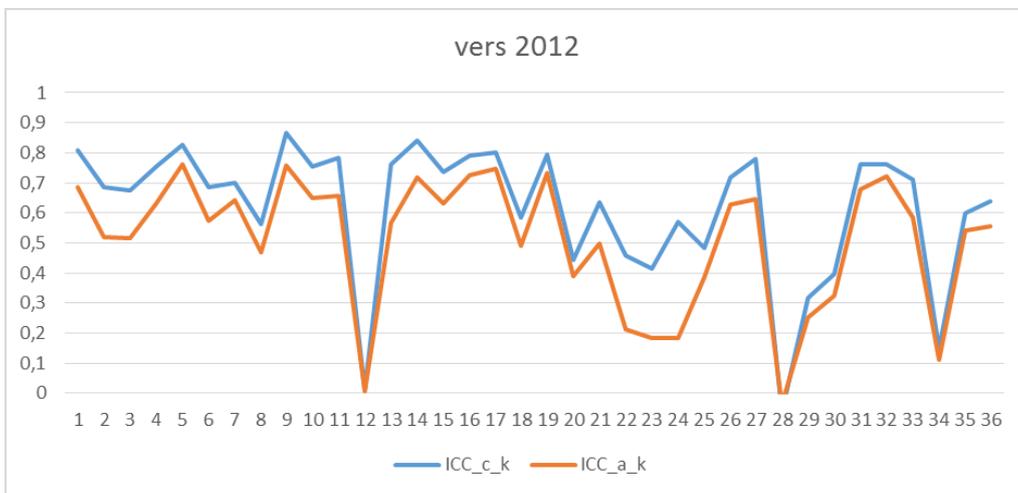


Figura 6. Resultados ICC conjunto de datos 2

Conclusiones

En nuestro trabajo hemos comprobado que hay una diferencia sensible en la fiabilidad de las evaluaciones realizadas por los alumnos de las presentaciones orales de sus compañeros. Puesto que entre ambas tomas de datos se han modificado no solo la rúbrica sino también el modo de administrarla (ver **Tabla 1**), no somos capaces de aislar cuál es el efecto de cada uno de los aspectos. Sin embargo, dado que la versión de la rúbrica de 2012 es, aparentemente, más adecuada y está avalada por un método de desarrollo sistemático, nos inclinamos a pensar en que el modo de administración de la rúbrica probablemente sí importa.

Como investigación futura planteamos superar alguna de las limitaciones de este trabajo, replicando el mismo procedimiento con las dos versiones de la rúbrica y comparando los resultados. También sería interesante comparar si la media de las puntuaciones de los alumnos es consistente con las puntuaciones otorgadas por los profesores. Por último, sería necesario investigar específicamente el efecto del tipo de control para intentar aumentar la responsabilidad de los alumnos en la tarea de evaluar. Existe cierta tendencia a pensar que los alumnos suelen estar poco motivados por aquellas actividades que no afectan directamente a su nota final de la asignatura. En este trabajo hemos planteado una alternativa en la que el profesor insistía y sensibilizaba en las ventajas de una correcta evaluación. Sin embargo, es probable que eso no haya sido suficiente para que las personas que evaluaban mostraran una implicación suficiente. En cualquier caso el diseño de la investigación presentado no permite dar una respuesta concluyente a esta cuestión.

Adicionalmente, podremos realizar un segundo análisis (

Figura 7) donde, 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 si hay algún criterio donde los evaluadores dan puntuaciones diferentes entre sí.

QUEST1	QUEST2	QUEST3	QUEST4	QUEST5	Rater	Grupc	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

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

Agradecimientos

Este trabajo ha sido realizado con la financiación de Universitat Politècnica de Valencia PIME/2015/A/009/A “Evaluación de los indicadores del comportamiento innovador en el alumno universitario”. Los autores agradecen los comentarios y sugerencias recibidas por los integrantes del grupo EICE-IEMA de la Universitat Politècnica de València

Referencias

- ANDREU, M. A. (2005). *Parrilla peer assessment oral presentation*.
- DE GREZ, L., VALCKE, M., & ROOZEN, I. (2012). How effective are self- and peer assessment of oral presentation skills compared with teachers assessments? *Active Learning in Higher Education*, 13(2), 129-142.
- DUNBAR, N. E., BROOKS, C. F., & KUBICKA-MILLER, T. (2006). Oral communication skills in higher education: Using a performance-based evaluation rubric to assess communication skills. *Innovative Higher Education*, 31(2), 115--128. doi:10.1007/s10755-006-9012-x
- EPIDAT. (2014). Manual epidat 4.1. Concordancia y consistencia. Santiago de Compostela: Servicio de Epidemiología de la Dirección General de Salud Pública de la Consellería de Sanidad (Xunta de Galicia)
- GARCÍA-ROS, R. (2011). Analysis and validation of a rubric to assess oral presentation skills in university contexts. *Electronic Journal of Research in Educational Psychology*, 9(3), 1043--1062.
- GARCÍA ROS, R., & Et Al. (2012). Designing and using rubrics in higher education: An innovation project in the psychology degree. *Electronic Journal of Research in Educational Psychology*, 10(3), 1478-1492.
- LANGAN, M., WHEATER, P., SHAW, E., HAINES, B., CULLEN, R., BOYLE, J., . . . PREZIOSI, R. (2005). Peer assessment of oral presentations: Effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment & Evaluation in Higher Education*, 30(1), 21-34.
- LEBRETON, J. M., & SENTER, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, 11(4), 815-852.
- MAGIN, D. J., & HELMORE, P. (2001). Peer and teacher assessments of oral presentation skills: How reliable are they? *Studies in Higher Education*, 26(3), 287-298.
- 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. (2016). Protocol: Inter-rater and intra-rater consistency validation of a rubric to assess oral presentation skills for university students. *Working Papers on Operations Management*, 7(1), (in press).
- MARIN-GARCIA, J. A., Aragonés Belgran, P., & Melón, G. (2014). Intra-rater and inter-rater consistency of pair wise comparison in evaluating the innovation competency for university students. *Working Papers on Operations Management*, 5(2), 24-46. doi:<http://dx.doi.org/10.4995/wpom.v5i2.3220>
- 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>.
- MORLEY, D. D. (2009). Spss macros for assessing the reliability and agreement of student evaluations of teaching. *Assessment & Evaluation in Higher Education*, 34(6), 659-671. doi:10.1080/02602930802474151
- VAN GINKEL, S., GULIKERS, J., BIEMANS, H., & MULDER, M. (2015). Towards a set of design principles for developing oral presentation competence: A synthesis of research in higher education. *Educational Research Review*, 14, 62--80. doi:10.1016/j.edurev.2015.02.002
- VILADRICH SEGUÉS, M. C., & DOVAL DIEGUEZ, E. (2011). *Medición: Fiabilidad y validez*. Bellaterra: Laboratori d'Estadística Aplicada i de Modelització (UAB).
- 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