Notas

Juan A. Marín-García
Los alumnos y los profesores como evaluadores.
Aplicación a la calificación de presentaciones orales.

Miguel Rumayor
Jacques Derrida: perspectiva y actualidad de la antimetafísica nietzscheana
en la educación para la ciudadanía.

Isabel Vilafranca Manguán y Mª Rosa Buxarrais Estrada
La educación para la ciudadanía en clave cosmopolita.
La propuesta de Martha Nussbaum.

Paloma Gavilán Bouzas
Aprendizaje cooperativo. Papel del conflicto sociocognitivo en el desarrollo
intelectual. Consecuencias pedagógicas.

Alfonso Osorio
Fundamentos filosóficos de la Psicología actual.

Los alumnos y los profesores como evaluadores. Aplicación a la calificación de presentaciones orales

por Juan A. MARÍN-GARCÍA
Universidad Politécnica de Valencia

Introducción

La educación universitaria se enfrenta en la actualidad a diferentes retos. Entre ellos, el papel que puede desempeñar la evaluación de los aprendizajes de los alumnos y la importancia del desarrollo de determinadas competencias profesionales (Dochy et al., 1999; Maclellan, 2001; Orsmond et al., 1996). Una forma de aunar estos dos aspectos es fomentar la participación de los alumnos en los procesos de evaluación del aprendizaje propio o el de sus compañeros. Sin embargo, la docencia tradicional no suele permitir esta posibilidad y la evaluación recae en los profesores, cuyas notas se consideran válidas e indiscutibles (Orsmond et al., 1996).

No obstante, cada vez es más habitual que se reconozca la importancia de formar a los alumnos universitarios para que desarrollen las habilidades necesarias para reflexionar criticamente sobre los resultados de su trabajo y el proceso seguido para completarlo (Bordas Alsina y Cabrera Rodriguez, 2001; Kwan y Leung, 1996). No sólo eso, sino que, además, sean capaces de evaluar también el trabajo de sus compañeros. Esta habilidad les será muy útil en su futuro trabajo profesional, sobre todo cuando la empresa en la que trabajen fomente el trabajo en equipo (Macpherson, 1999).

La investigación sobre la evaluación de los compañeros suele tener como objeto la puntuación de ensayos, proyectos de grupo o presentaciones orales (Dochy et al., 1999; Magin, 2001b) y, en algunos casos, posters (Orsmond et al., 2000). Como se han comparado actividades muy diferentes, no es de extrañar que existan resultados divergentes.

El número de investigaciones realizadas hasta la fecha no es demasiado eleva-

do y sería recomendable ampliar las experiencias relacionadas con las evaluaciones de presentaciones orales (MacAlpine, 1999) en diferentes disciplinas (Gatfield, 1999; Kwan y Leung, 1996). Además, parece necesario incorporar nuevas formas de analizar los datos, que sean más robustas y que permitan integrar las conclusiones de diferentes investigaciones (Kwan y Leung, 1996; Magin, 2001a).

Nuestro trabajo de campo se ha centrado en comprobar la validez de las puntuaciones de los alumnos, comparándolas con las del profesor de la asignatura. Pero también hemos comprobado si las puntuaciones del profesor de la asignatura son válidas. Para ello, las hemos comparado con las puntuaciones de 4 profesores más. Además analizaremos la fiabilidad de las puntuaciones cuando contamos con varios evaluadores para cada presentación, tanto si son alumnos como si son profesores.

Para ello, hemos analizado las formas de incorporar a los estudiantes en el proceso de evaluación y las conclusiones a las que ha llegado la investigación previa. Sólo hemos tenido en cuenta los estudios relacionados con la evaluación de presentaciones orales, que será la actividad realizada en el trabajo de campo. En el apartado de introducción, podremos apreciar que uno de los aspectos mas conflictivos tiene que ver con la validez/fiabilidad de las puntuaciones de los alumnos. También resumiremos las recomendaciones que se han propuesto para mejorar el proceso de participación de los alumnos en la evaluación. Concluiremos la revisión de la literatura poniendo de manifiesto qué análisis estadísticos se han utilizado para comprobar la fiabilidad de las puntuaciones realizadas por los alumnos, los resultados obtenidos en otras investigaciones y las limitaciones de los mismos.

Evaluacion por parte de los compañeros *(peer assessment)*

La evaluación por parte de los compañeros, *"peer assessment"*, (PA) consiste en un proceso a través del cual un grupo de personas puntúa a sus compañeros (Dochy et al., 1999).

Varios autores (Dochy et al., 1999; Kwan y Leung, 1996) resumen algunas de las ventajas del PA: los alumnos tienen más confianza en sus habilidades; mejora la percepción del alumno sobre la calidad de su trabajo; los alumnos reflexionan más sobre su conducta y sus resultados; mejora los resultados de los alumnos en los exámenes; mejora la calidad y la eficacia del aprendizaje; los alumnos toman más responsabilidades en el proceso de aprendizaje y aumenta la satisfacción de los alumnos. Por otra parte, se considera que desarrollar en los estudiantes la habilidad de evaluar el trabajo propio o el de los compañeros, es un elemento importante en el proceso de aprendizaje (Kwan y Leung, 1996, Reynolds y Trehan, 2000). De hecho, es una de las competencias profesionales para las que debemos formar a nuestros universitarios si queremos formar profesionales reflexivos y fomentar el aprendizaje a lo "largo de toda la vida" (Cheng y Warren,1999; Dochy et al.,

1999; MacAlpine, 1999; Macpherson, 1999).

Al revisar las investigaciones precedentes, los principales problemas que se citan acerca de validez y fiabilidad del PA son:

- Se puede pactar la nota o puntuar mejor a los amigos o a uno mismo que a los demás (Dochy et al., 1999; Magin, 2001b; Magin y Helmore, 2001).
- Los alumnos no suelen usar todo el rango de la escala disponible y tienen a concentrar sus puntuaciones en el tramo medio (Dochy et al., 1999; Macpherson, 1999).
- Los alumnos no tienen el mismo nivel de referencia que los profesores (Magin y Helmore, 2001).
- Los alumnos son malos evaluadores de la actividad sujeta a puntuación y sus puntuaciones no se parecen a las de los profesores (Dochy et al., 1999; Macpherson, 1999; Magin y Helmore, 2001).

Las resistencias a incorporar las puntuaciones de los alumnos como evaluación sumativa, debidas a los problemas mencionados, no son sólo de los profesores. Algunos estudiantes desconfían de la calidad de las puntuaciones de sus compañeros o de ellos mismos y no se sienten a gusto con el sistema (Kwan y Leung, 1996; Macpherson, 1999).

Por otra parte, hemos recogido las recomendaciones que hacen los diferentes autores para mejorar el proceso de PA. En este sentido, al poner en marcha estos sistemas es aconsejable lo siguiente:

- La evaluación es una habilidad que se puede mejorar con la repetición y el entrenamiento (Dochy et al., 1999; Kwan y Leung, 1996; Macpherson, 1999; Magin, 2001 a; Searby y Ewers, 1997).
- Es necesario reservar tiempo para que el alumno pueda realizar sus puntuaciones. También es posible que los alumnos necesiten apoyo u orientaciones en algún momento del proceso (Dochy et al., 1999; Magin, 2001a).
- Si queremos usar PA como herramienta formativa, debemos ayudar a que los estudiantes vean estas actividades como una herramienta para facilitar su aprendizaje (Dochy et al., 1999).
- Usar los mismos criterios para todos los evaluadores, que estén establecidos de antemano, sean conocidos por los alumnos y estén formulados de una manera clara y sencilla. Si es posible, se recomienda negociarlos con los estudiantes (Cheng y Warren, 1999; Dochy et al., 1999; Kwan y Leung, 1996; MacAlpine, 1999; Macpherson, 1999; Magin, 2001a; Orsmond et al., 2000; Searby y Ewers, 1997; Sullivan y Hall, 1997).

- Aunque las puntuaciones individuales de una persona no sean fiables, cuando se promedian las puntuaciones de varias personas, la fiabilidad es muy elevada (Magin y Helmore, 2001).

- Usar PA como una parte de la nota, que complementa la nota entregada por los profesores (Dochy et al., 1999; Kwan y Leung, 1996; Magin, 2001a; Searby y Ewers, 1997).

En definitiva, parece que el principal punto de desacuerdo entre las investigaciones es el grado de fiabilidad de las puntuaciones de los alumnos. Si se mejorara este aspecto, permitiría reducir algunas de las resistencias que hay por parte de los profesores y de los alumnos para adoptar estos sistemas. Por eso, el objeto de nuestro trabajo será comprobar el grado de fiabilidad que se obtiene en una implantación concreta de PA (evaluación de presentaciones orales), cuando se tienen en cuenta las recomendaciones realizadas por la literatura científica sobre el tema.

Validez y fiabilidad de las puntuaciones de los alumnos en presentaciones orales

Se han utilizado diferentes medidas para establecer el grado de acuerdo entre las puntuaciones de los alumnos y las de los profesores. La más habitual es usar el coeficiente de correlación entre la media de las puntuaciones de los alumnos con la puntuación del profesor (o la media de las puntuaciones de los profesores, si hay varios puntuando) (Al-Fallay, 2004; Cheng y Warren, 1999; Falchikov y Goldfinch, 2000; Freeman, 1995; Kwan y Leung, 1996; Langan et al., 2005; MacAlpine, 1999, Macpherson, 1999; Magin y Helmore, 2001). Otras medidas utilizadas son el porcentaje de alumnos que dan una nota comprendida en un intervalo de confianza sobre la nota del profesor (normalmente una desviación estandar) (Falchikov, 1995; Freeman, 1995; Kwan y Leung, 1996), comparar la varianza de las notas de los alumnos y de los profesores o hacer un T-test para la diferencia de la medias entre las notas de los alumnos y las de los profesores (Cheng y Warren, 1999; Freeman, 1995; Kwan y Leung, 1996; Magin y Helmore, 2001; Ward et al., 2002). Recientemente, algunos estudios utilizan una medida asociada al análisis de la varianza (ANOVA) para determinar la la fiabilidad entre evaluadores (Magin, 2001a).

Podemos comprobar que el grado de acuerdo entre las puntuaciones promedio de los profesores y las del grupo de alumnos, es bastante elevado cuando se evalúan presentaciones orales.

Por un lado, las correlaciones entre las puntuaciones de los alumnos y las de los profesores son moderadas o altas. Los valores se sitúan entre 0,44 y 0,79 en 3 de los estudios citados (Falchikov y Goldfinch, 2000) y en los trabajos de varios autores (Cheng y Warren, 1999; Freeman, 1995; Kwan y Leung, 1996; Macpherson, 1999; Magin y Helmore,

2001). Mientras que son mayores que 0,80 en otros 3 estudios citados (Falchikov y Goldfinch, 2000) y en otras dos investigaciones (Al-Fallay, 2004; Langan et al., 2005; MacAlpine, 1999).

Por otro lado, el grado de coincidencia de las notas de los alumnos y las del profesor es muy elevado 98% (Falchikov, 1995), 95% (Freeman, 1995) y 70% (Kwan y Leung, 1996). También se puede confirmar que no aparecen diferencias significativas entre notas de alumnos y profesores al aplicar una prueba T (Freeman, 1995; Kwan y Leung, 1996). Sin embargo, la dispersión de las notas de los alumnos es sensiblemente menor que las de los profesores (Cheng y Warren, 1999; Freeman, 1995; Kwan y Leung, 1996; Magin y Helmore, 2001). En otras palabras, que los alumnos tienden a concentrar sus notas y discriminan menos que los profesores.

No obstante, debemos tener presentes una serie de advertencias realizadas por Ward y colaboradores (2002). En primer lugar, casi todos los estudios consideran que la nota del profesor es correcta y que la diferencia de notas con los alumnos es debida a que los alumnos no son tan buenos puntuando como los profesores. No obstante, esta diferencia también puede ser debida a que las puntuaciones del profesor no sean tan válidas y fiables como cabría pensar (Falchikov y Goldfinch, 2000; Magin y Helmore, 2001; Orsmond et al., 1996; Pascual Gomez y Gaviria Soto, 2004). La validez tiene que ver con puntuar lo que se pretende medir y, al mismo tiempo, que lo que se pretenda medir sea relevante. Es decir, que sea representativo de los conocimientos/habilidades, perseguidas en la asignatura, que el alumno ha adquirido. Las sugerencias para superar este problema es mejorar la fiabilidad de las puntuaciones de los profesores usando el promedio de las puntuaciones de varios profesores expertos (Ward et al., 2002).

En segundo lugar, aunque existan una serie de criterios para puntuar, no hay garantía de que todos los evaluadores los interpreten de la misma manera. La mejor forma de evitar este problema es proporcionar varios criterios, lo más sencillos posibles e incluyendo guías explícitas en la parrilla de calificación (Ward et al., 2002). Otra posibilidad es tipificar las puntuaciones de los alumnos o de los profesores antes de promediarlas.

Por ultimo, Ward y colaboradores (2002) también comentan un posible problema cuando sólo se analizan los datos a nivel de grupo. Es decir, cuando comparamos la media de las puntuaciones de un grupo de evaluadores con la nota de un experto. Este tipo de análisis no nos indica el grado de acuerdo de cada alumno individual con la nota del profesor y no nos aporta información relevante cuando sólo se dispone de una nota originada por los alumnos (bien porque se autoevalúan o bien porque cada trabajo es evaluado sólo por un compañero).

En este artículo, pretendemos utilizar las dos primeras recomendaciones de

Ward y colaboradores (2002) y dejaremos para una investigación posterior la evaluación de los diferentes niveles de análisis.

Objetivos y metodología

Teniendo en cuenta las investigaciones que hemos comentado en los apartados anteriores, nos parece interesante continuar aportando datos para clarificar hasta qué punto podemos confiar en las puntuaciones que proporcionan los alumnos. De este modo, pretendemos reflexionar sobre la posibilidad de incorporarlas a la evaluación sumativa de nuestras asignaturas.

Las preguntas que pretendemos responder con nuestra investigación son las siguientes:

1. ¿Son válidas las puntuaciones de un sólo profesor? Es decir, ¿concuerdan con las de un grupo de profesores?

2. ¿Es válido el promedio de las puntuaciones de varios alumnos evaluando una misma presentación? Es decir, ¿concuerda con la puntuación del profesor?

3. ¿Qué grado de fiabilidad tienen las puntuaciones de un sólo profesor si las comparamos con la de un grupo de profesores?

4. ¿Son fiables las puntuaciones de los alumnos? ¿Cuántos alumnos evaluadores serían necesarios para conseguir una fiabilidad similar a la de un profesor?

Para dar respuesta a las preguntas de investigación, utilizaremos los procedimientos siguientes:

- La validez de las puntuaciones de un profesor la mediremos contrastando los puntos entregados por el profesor de la asignatura con la media de las puntuaciones de los otros 4 profesores que participan en esta investigación. Como no fue posible que asistieran más profesores a la sesión de exposición de final de curso, utilizaremos los datos provenientes de la puntuación de las presentaciones grabadas en vídeo, tanto para el profesor de la asignatura como para los profesores colaboradores en la investigación. De este modo, la situación es común para todos ellos. Con estos datos calcularemos la correlación y haremos una prueba de T de diferencia de medias para comprobar si las puntuaciones se corresponden a una misma población. También calcularemos cuántas puntuaciones del profesor de la asignatura se diferencian de las del promedio de los otros 4 profesores, en menos de una desviación estándar de las puntuaciones del grupo de profesores (Cheng y Warren, 1999; Freeman, 1995; Kwan y Leung, 1996; Magin y Helmore, 2001).

- La validez del promedio de las puntuaciones de los alumnos la intentaremos probar con un pro-

cedimiento análogo. Pero en este caso, el valor de referencia será las puntuaciones otorgadas por el profesor de la asignatura el día de las presentaciones. Por lo tanto, la situación en la que se puntuaba las presentaciones es la misma para ambos conjuntos de datos.

- La fiabilidad de las puntuaciones de los profesores y de los alumnos las mediremos utilizando el procedimiento desarrollado por Magin y Helmore (2001) (ver apéndice A). Nuestra hipótesis de partida es que los profesores tendrán un grado mayor de fiabilidad que los alumnos. El motivo para ello no será la experiencia en la materia que se imparte, puesto que la mayoría de los profesores que participan en la investigación imparten asignaturas muy diferentes a la evaluada, sino que la fiabilidad de los profesores se supone asociada con el hecho de estar habituados a puntuar y a discriminar las puntuaciones de sus alumnos.

- Por último completaremos los análisis con la estimación del número de alumnos o de profesores que deberían evaluar simultáneamente una exposición para que su fiabilidad fuese similar a la de la puntuación de un sólo profesor experto (tanto en la materia impartida como en evaluar presentaciones orales). Para ello, seguiremos el procedimiento sugerido por Magin (2001a) (ver apéndice A).

Kane y Lawler (1978), sugieren tres procedimientos para la puntuación por parte de los compañeros: rango ordenado, nominación y puntuación. Nosotros hemos elegido el tercero de ellos, donde cada persona puntúa a sus compañeros de acuerdo con su rendimiento, utilizando una escala de puntuación. Consideramos que es el de más fácil aplicación (aunque los otros métodos puedan ser más fiables o permitir un mayor grado de discriminación).

Para fomentar una mayor implicación de los alumnos, decidimos que los criterios a evaluar fuesen seleccionados por los estudiantes, tal como describiremos más adelante. 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 están 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 asignatura (gestión de empresas) se imparte durante 15 semanas lectivas en clases de 2,5 horas semanales. 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 tra-

bajo ante los demás compañeros en clase. Esta exposición sería grabada en vídeo. 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 23 presentaciones y en ellas intervinieron 44 alumnos.

Un mes antes de la exposición se explicó la actividad durante una sesión de clase. Una vez aclaradas las dudas generales se formaron dos grupos de 15-20 alumnos, que se reunieron para seleccionar los criterios que se utilizarían para puntuar la actividad. Se empezó con una "tormenta de ideas con grupo nominal" sobre los aspectos que definen una buena exposición oral. Posteriormente se filtraron los criterios (utilizando la técnica de diagrama de afinidad) teniendo como premisa que se seleccionaran sólo aquellos que fuesen fáciles de comprender y de observar objetivamente por los alumnos que actuarían como evaluadores. Tanto la tormenta de ideas como el diagrama de afinidad son dos técnicas contenidas en el programa de la asignatura por lo que estas reuniones sirvieron para practicarlas. El profesor se encargó de recoger las dos listas y sintetizarlas después de la reunión. También incluyó algunas guías para facilitar la puntuación.

Tres semanas antes de la exposición, durante la clase semanal, se presentó la lista definitiva a los alumnos (ver apéndice B) que la utilizaron para evaluar a 3 compañeros suyos que actuaban como portavoces de una de las actividades de grupo realizadas durante la sesión. Además, se puso a disposición de los alumnos, en la WEB, las grabaciones en vídeo de las presentaciones realizadas durante la asignatura. Estas presentaciones eran independientes de la actividad objetivo de este artículo y consistían en la actuación de diferentes alumnos (casi el 50% de los asistentes) como portavoces de las actividades de grupos que se iban realizando durante las sesiones de clase. El objetivo era que, durante la semana, los alumnos se evaluasen a sí mismos y a otros dos compañeros utilizando la plantilla. De este modo, recibían *feedback* de sus habilidades como oradores y se familiarizaban con el uso del instrumento. En la clase siguiente se dio la oportunidad de que los alumnos comentasen las dificultades surgidas durante la utilización de la parrilla. No fue necesario realizar ninguna aclaración ni rectificación de la misma.

Las dos semanas siguientes, los alumnos trabajaron por su cuenta realizando las entrevistas. El día de la presentación, los alumnos contaban con una hora de clase para preparar sus presentaciones (de 3 minutos cada una como máximo). Durante la segunda hora de clase se procedió a realizar las presentaciones que fueron puntuadas por sus compañeros y por el profesor.

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. El profesor evaluó todas las presentaciones, usando para ello la misma parrilla de puntuación que los alumnos.

Cuatro meses después de la exposición, un grupo de cuatro profesores se reunió con el profesor de la asignatura para evaluar las presentaciones orales, que habían sido grabadas en video. De los cuatro profesores, uno impartía la misma asignatura, pero en un centro distinto; otro pertenecía al mismo departamento y los otros dos pertenecían a un departamento diferente.

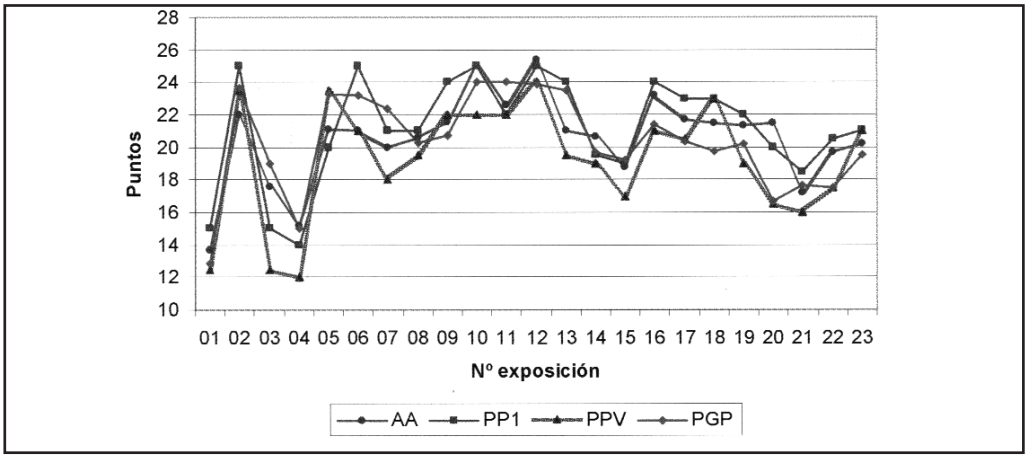
En total recogimos 4 conjuntos de datos:

- Las puntuaciones que dan los alumnos a las presentaciones de sus compañeros (cada exposición es evaluada por 7-12 alumnos) (AA).
- Las puntuaciones del profesor de la asignatura el día de la exposición (PP1).
- Las puntuaciones del profesor de la asignatura viendo el vídeo de las presentaciones (PPV).
- Las puntuaciones de cuatro profesores viendo el vídeo de las presentaciones (PGP).

Análisis y discusión de los resultados

En la Figura 1 presentamos el gráfico de las puntuaciones obtenidas por cada una de las presentaciones realizadas. Hemos calculado el promedio en aquellas series de datos donde se disponía de más de

FIGURA 1: *Puntos obtenidos en las presentaciones.*



AA: promedio de las puntuaciones de los alumnos el día de la exposición.
PP1: puntuación del profesor de la asignatura el día de la exposición.
PPV: puntuación del profesor de la asignatura viendo el vídeo.
PGP: promedio de las puntuaciones de 4 profesores viendo vídeo.

TABLA 1: *Estadísticos descriptivos.*

	N	Media	Desviación típica	Mínimo	Máximo	AA	PP1	PPV	PGP
AA: promedio de las puntuaciones de los alumnos que evalúan esa presentación	242	20,538	2,7155	13,7	25,4		,875(**)	,856(**)	,811(**)
PP1: puntuaciones del profesor de la asignatura el día de la presentación	23	21,152	3,2768	14,0	25,0	-	1	,878(**)	,791(**)
PPV: puntuaciones del profesor de la asignatura sobre video	23	19,239	3,5543	12,0	24,0	-	-	1	,808(**)
PGP: promedio de las puntuaciones de los otros 4 profesores sobre video	87	20,326	3,0168	12,8	24,0	-	-	-	1

** La correlación es significativa al nivel 0,01 (bilateral).

una puntuación por exposición (AA y PGP). A simple vista se puede observar que existe bastante similitud entre las diferentes fuentes de datos. No obstante completaremos una serie de pruebas estadísticas para dar rigor a esta interpretación.

En primer lugar, hemos calculado los estadísticos descriptivos y hemos comprobado que las cuatro series de datos (AA, PP1, PPV y PGP) tienen una distribución normal, mediante la Prueba de Kolmogorov-Smirnov para una muestra (VVAA, 2002). También resumimos en la Tabla 1 las correlaciones de Pearson entre las variables.

Las puntuaciones realizadas por el profesor de la asignatura de las presentaciones grabadas en vídeo (PPV) están altamente correlacionadas con el promedio de las puntuaciones de los otros cuatro profesores (PGP) (0,808). Las diferencias entre las medias de ambas puntuaciones son significativas con una prueba T (a=0,02). Esto indica que, en promedio, las puntuaciones del profesor de la asignatura son 1,08 puntos menores que las de los otros profesores. En total hay 4 presentaciones (17° o) en las que la diferencia entre las notas de la serie PPV y las de la serie PGP son mayores que 3 puntos (una desviación estándar de la serie PGP).

De manera análoga, la correlación entre el promedio de las puntuaciones de los alumnos (AA) y la del profesor de la asignatura, el día de las presentaciones (PP1) es significativa y muy elevada (0,875). Las diferencias entre las medias de ambas puntuaciones no son significativas con una prueba T. En total, sólo hay una exposición (4%) en la que la puntuación AA se diferencia de PP1 en más de una desviación estandar de las notas PP1 (3,27 puntos).

Por lo tanto, respecto a las dos primeras preguntas de nuestra investigación podemos concluir que las puntuaciones de un único profesor son razonablemente válidas, pues coinciden bastante con las de un grupo de cuatro profesores. La diferencia promedio entre ambas puntuaciones, aunque es significativa estadística-

mente, sólo es de 1 punto sobre un total de 27. Por otra parte, en el 83% de los casos la diferencia entre las puntuaciones PPV y PGP es menor que 3 puntos sobre una escala de 27 puntos. No disponemos de datos de investigaciones similares que hayan estudiado estos aspectos, por lo que no podemos valorarlo en comparación con otros hallazgos.

Del mismo modo, y de manera más remarcada, son válidos los promedios de las puntuaciones realizadas por los alumnos a sus compañeros. De hecho, la concordancia de estos promedios con la nota del profesor de la asignatura es prácticamente total. Nuestros resultados coinciden con los estudios anteriores que encontraron altas correlaciones entre las puntuaciones de los alumnos y las de los profesores (Al-Fallay, 2004; Falchikov, 1995; Falchikov y Goldfinch, 2000; Freeman, 1995; Langan et al., 2005; MacAlpine, 1999; Macpherson, 1999; Magin y Helmore, 2001). Nuestros datos también coinciden con todos los estudios que han observado que la dispersión de las puntuaciones de los alumnos es menor y que discriminan menos (Cheng Warren, 1999; Falchikov y Goldfinch, 2000; Freeman, 1995; Kwan y Leung, 1996); Magin y Helmore, 2001). Aunque en nuestro caso, las diferencias de dispersión son bastante pequeñas. Por último, las puntuaciones que dan los alumnos a sus compañeros son ligeramente menores que las que da el profesor.

Una vez analizadas las preguntas relacionadas con la validez, pasaremos a comentar los resultados obtenidos con la fiabilidad de las puntuaciones de los alumnos y del profesor. En ambos casos seguiremos un procedimiento análogo.

En primer lugar, calcularemos la fiabilidad entre evaluadores (r_{nn}) del grupo de alumnos que puntúan a sus compañeros o del grupo de profesores que ven el vídeo de las presentaciones, sin incluir la puntuación del profesor de la asignatura. Posteriormente, añadiremos a los análisis las puntuaciones del profesor de la asignatura para estimar cuál sería su fiabilidad (r^*_{nn}). Por último, si la fiabilidad estimada del profesor de la asignatura (r_{tt}) es superior a la de los alumnos o la del conjunto de profesores colaboradores en la investigación (r_{ll}), calcularemos cuántos alumnos o profesores de otras disciplinas (DeltaN) sería necesario que evaluaran cada exposición para que el promedio de sus notas tuviese una fiabilidad similar a la del profesor de la asignatura.

Los resultados de estos análisis (Tabla 2) no pueden ser más elocuentes. La fiabilidad de las puntuaciones de los alumnos (r_{nn}=0,90) es muy elevada cuando se incorpora una cantidad grande de evaluadores (en nuestro caso más de 10 por exposición). Sin embargo, la fiabilidad estimada de las puntuaciones de un único alumno evaluador es un poco baja (r_{ll}=0,47), pero prácticamente igual a la de uno sólo de los 4 profesores que puntuaban los vídeos.

Ahora bien, la fiabilidad estimada de las puntuaciones del profesor de la asig-

TABLA 2: *Análisis de la fiabilidad de las puntuaciones de los alumnos y profesores.*

	Puntuación de los alumnos a los compañeros el día de la exposición	Puntuación del grupo de profesores viendo el vídeo de las presentaciones
Número de evaluadores	43	4
Numero de presentaciones evaluadas	23	23
Numero de observaciones	242	87
Promedio de evaluadores por exposición (N)	10,52	3,78
Estadístico ANOVA de las puntuaciones de n evaluadores (F)	10,34	4,20
Estadístico ANOVA cuando se añaden las puntuaciones del profesor de la asignatura a las puntuaciones de los n evaluadores (F*)	12,24	6,13
Fiabilidad de las puntuaciones de N evaluadores (r_{nn})	0,90	0,76
Estimación de la fiabilidad de un solo evaluador (r_{11})	0,47	0,46
Fiabilidad de las puntuaciones cuando se incorporan las del profesor de la asignatura a los n evaluadores (r^*_{nn})	0,92	0,84
Estimación de la fiabilidad de las puntuaciones del profesor de la asignatura (r_{tt})	0,66	0,66
Estimación del número de alumnos/profesores que deberían puntuar cada exposición para que su fiabilidad fuese similar a la del profesor de la asignatura (DeltaN)	2,14	2,27

natura son sensiblemente mayores (r_{tt}=0,66) y prácticamente constantes al contrastarla con ambas series de datos. Tanto la fiabilidad de las puntuaciones de los alumnos como la del profesor de la asignatura (r_{nn}, r_{ll} y r_{tt}) son más elevadas que las obtenidas en la investigación de Magin y Helmore (2001). Sin embargo, la fiabilidad de las puntuaciones de los otros cuatro profesores es prácticamente igual a la obtenida en esa investigación.

Consideramos que puede haber varios motivos por los cuales la fiabilidad sea tan alta. En primer lugar, el hecho de que sólo hemos incluido criterios bastante objetivos, y que cada uno de los criterios tenía unas guías definidas para la puntuación. Somos conscientes de que hemos dejado de lado aspectos subjetivos que también son importantes en la evaluación de presentaciones en público (por ejemplo, si gusta, si capta la atención del oyente, si interesa...). Sin embargo, nos interesaba más fomentar la fiabilidad de las puntuaciones que contemplar los aspectos subjetivos. Además, entendemos que cumplir con los aspectos objetivos observados era una condición necesaria para una buena presentación, es decir, si se

cumplen no garantizan una buena presentación, pero si no se cumplen seguro que la presentación no puede ser exitosa. Por otra parte, la parrilla la planteamos como un instrumento formativo ya que servía como guía a los alumnos, indicando qué debían hacer durante la exposición (por ejemplo, hablar alto y claro, contacto visual con la audiencia...).

En segundo lugar, tanto los alumnos como el profesor de la asignatura han podido probar previamente la parrilla y familiarizarse con los criterios y la forma de puntuación antes del día de las presentaciones evaluadas en este artículo. Sin embargo, los otros cuatro profesores que colaboraron en la investigación no disfrutaron de este beneficio. Ellos conocieron los criterios en la misma sesión que puntuaron los vídeos de las presentaciones. Sólo recibieron una explicación de 15 minutos por parte del profesor de la asignatura, tras la cual tuvieron unos minutos para leer los criterios y consultar dudas. Una vez resueltas sus dudas, empezó la proyección del vídeo y puntuaron las presentaciones. Ésta puede ser una de las causas por la que la fiabilidad estimada de uno de estos profesores (r_{ll}) es sensiblemente menor a la del profesor de la asignatura (r_{tt}). Consideramos que estos cuatro profesores tenían una experiencia similar, y en algunos casos mucho mayor, que el profesor de la asignatura en la evaluación de presentaciones orales de sus alumnos. Por lo tanto, esa mejora de fiabilidad sólo puede explicarse por la familiaridad con los criterios utilizados. Por otra parte, puesto que r_{ll} es similar en los profesores (personas habituadas a puntuar a sus alumnos, pero no familiarizados con la parrilla de criterios concreta que hemos utilizado en esta asignatura) y en los alumnos (personas no acostumbradas a evaluar a sus compañeros, pero entrenados en el uso de la parrilla de criterios), podemos intuir que los criterios utilizados y el entrenamiento previo han servido como sustituto de la experiencia.

Conclusiones

En primer lugar, queremos advertir que las conclusiones que vamos a extraer sólo son aplicables a situaciones similares a la relatada: presentaciones orales, evaluadas con parrillas construídas por los propios alumnos que han sido entrenados en su uso. Además, las puntuaciones de los alumnos contribuían a la evaluación sumativa de sus compañeros, pero tambien recibían una nota como evaluadores, lo que les obligaba a tomarse en serio el proceso. También intentamos motivar a los alumnos comentándoles que en su futuro trabajo como mandos es probable que tuviesen que evaluar el rendimiento de sus subordinados o, incluso, de sus compañeros. Todas estas características se han extraído de las recomendaciones de investigaciones previas y sin ellas es probable que los resultados fuesen diferentes.

En estas condiciones, podemos considerar que las puntuaciones de las presentaciones, calculadas como el promedio de los puntos recibidos de varios alumnos actuando como evaluadores, no se diferencian de manera significativa de las puntuaciones que otorga el profesor de la

asignatura. Por ello, podemos utilizar las notas derivadas del PA en la evaluación final de los alumnos sin que ello distorsione los resultados.

Es más, las puntuaciones del profesor de la asignatura presentan algunas diferencias al compararlas con el promedio de notas que propone el grupo de 4 profesores que ha colaborado en la investigación. Aunque esas diferencias no son grandes, nos recuerdan que, por muy experimentado que sea un profesor, no podemos dar por sentado que las puntuaciones que realiza son absolutamente acertadas.

Esta conclusión es corroborada por el dato de que las fiabilidades estimadas de un solo evaluador son moderadamente bajas en el caso del profesor de la asignatura y algo bajas en el caso de los alumnos o de los otros profesores participantes. Por lo tanto, lo más aconsejable, desde el punto de vista de la fiabilidad, sería utilizar las puntuaciones de varios evaluadores simultáneamente.

Los resultados de nuestro trabajo parecen indicar con bastante claridad que los alumnos pueden llegar a ser unos buenos evaluadores de las presentaciones orales de sus compañeros y que sus puntuaciones podrían servir para establecer las notas de sus compañeros sin que se aprecien diferencias con las notas que propone el profesor de la asignatura.

Si podemos fiarnos de las puntuaciones de los alumnos, los profesores podemos descargarnos de parte de la responsabilidad y tiempo dedicado a evaluación. De manera que, al mismo tiempo que fomentamos la autonomía, responsabilidad y participación de los alumnos en el proceso (aspectos que suelen originar una mayor motivación de los estudiantes), conseguimos liberar tiempo en las ya de por sí sobrecargadas agendas de los profesores.

Si no queremos dejar toda la responsabilidad en manos de los alumnos, tenemos dos posibilidades: hacer intervenir a varios profesores o utilizar CA. En el contexto de esta asignatura lo más habitual es no poder contar con la participación de más profesores (bien porque es impartida por un sólo profesor o porque, aún habiendo varios profesores, no dispongan de tiempo para invertirlo en duplicar las evaluaciones). Por lo tanto, se podría mejorar la fiabilidad si cada exposición fuese puntuada por el profesor de la asignatura junto con varios alumnos (mínimo entre dos y cuatro). Con ello podríamos tener unos valores de r^*_{nn} superiores a 0,80 que puede considerarse bastante adecuados (Magin y Helmore, 2001). Con este procedimiento, el profesor no se ahorraría tiempo, pues tiene que puntuar las exposiciones y además invertir tiempo en promediar las puntuaciones de los alumnos participantes. Pero se mejoraría la fiabilidad de la evaluación y se conseguiría implicar a los alumnos en el proceso, con los beneficios formativos y de motivación que ello conllevaría.

También queremos recalcar, como factores importantes para fomentar una ele-

vada fiabilidad en las puntuaciones, una selección de criterios lo más objetivos posibles y la creación de guías de puntuación, la familiaridad o entrenamiento con los criterios de evaluación y la experiencia en la evaluación de presentaciones orales. Cuanto más presentes estén estos factores, mayor será la fiabilidad de las puntuaciones. Es más, manejando adecuadamente los dos factores que caen bajo el campo de actuación de un profesor (selección de criterios y entrenamiento) se pueden conseguir fiabilidades muy elevadas, incluso con pocos evaluadores de cada presentación. Por lo tanto, no parece haber limitación, al menos por el aspecto de fiabilidad y validez, para fomentar una mayor participación de los alumnos a la puntuación de nota final de sus compañeros.

La forma en que se ha construido la parrilla de criterios para la puntuación, con la participación activa de los alumnos, y las actividades para que los alumnos se entrenaran en el uso de la parrilla, también puede haber influido favorablemente en la fiabilidad de las puntuaciones. No obstante, no estamos en condición de valorar su efecto en estos momentos. Nuestra intención es continuar la investigación con alumnos de cursos posteriores que no han participado en la elaboración de la parrilla y ver los resultados que se producen. También nos gustaría analizar qué pasa cuando los alumnos usan la parrilla en las mismas condiciones que el grupo de 4 profesores. Es decir, viendo el vídeo de las presentaciones tras una explicación de los criterios durante 15 minutos, sin entrenarse con el uso de la parrilla.

Por último, debemos tener en cuenta que la implantación del PA y su utilización en la nota final de los alumnos, se verá dificultada en entornos universitarios tradicionales. Sobre todo, si la transferencia de responsabilidades cambia el equilibrio de poder entre los profesores y los alumnos (Dochy et al., 1999; Macpherson, 1999; Tariq et al., 1998). Además, cuando se pone en marcha PA, la gestión del proceso se complica y se invierte más tiempo. Por lo tanto, se recomienda revisar, simplificar y computerizar el procedimiento para no sobrecargar a los docentes (Cheng y Warren, 1999).

Dirección del autor: Juan A. Marín-García. Departamento de Organización de Empresas. Universidad Politécnica de Valencia. Edificio 7D. App. Correos 22012 46071 Valencia. SPAIN. Correo Electrónico: jmarin@omp.upv.es

Fecha de recepción de la versión definitiva de este artículo: 1.VI.2008.

Bibliografía

AL-FALLAY, I. (2004) The Role of Some Selected Psychological and Personality Traits of the Rater in the Accuracy of Self-and Peer-Assessment, *System,* 32:3, pp. 407-427.

BORDAS ALSINA, I. y CABRERA RODRIGUEZ, F. A. (2001) Estrategias de evaluación de los aprendizajes centradas en el proceso, **revista española de pedagogía**, 59:218, pp 25-48.

CHENG, W. y WARREN. M. (1999) Peer and Teacher Assessment of the Oral and Written Tasks of a Group Project, *Assesment & Evaluation in Higher Education,* 24:3, pp. 301-314.

DOCHY. F.; SEGERS, M. y SLUIJSMANS, D. (1999) The Use of Self-, Peer and Co-Assessment in Higher Education: a Review, *Studies in Higher Education,* 24:3, pp. 331-350.

FALCHIKOV, N. (1995) Peer Feedback Marking - Developing Peer Assessment, *Innovations in Education and Training International,* 32:2, pp. 175-187.

FALCHIKOV, N. y GOLDFINCH, J. (2000) Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks, *Review of Educational Research,* 70:3, pp. 287-322.

FREEMAN, M. (1995) Peer Assessment by Groups of Group Work, *Assesment & Evaluation in Higher Education,* 20, pp. 289-300.

GATFIELD, T. (1999) Examining Student Satisfaction With Group Projects and Peer Assessment, *Assesment & Evaluation in Higher Education.* 24:4, pp. 365-377.

KANE, J. y I.AWLER III, E. E. (1978) Methods of Peer Assessment, *Psychological Bulletin,* 85:3, pp. 555-586.

KWAN, K.-P. y LEUNG, R. (1996) Tutor Versus Peer Group Assessment of Student Performance in a Simulation Training Exercise, *Assesment & Evaluation in Higher Education,* 21:3, pp. 205-214.

LANGAN, M.; WHEATER, P.; SHAW, E.; HAINES, B.; CULLEN, R.; BOYLE, J.; PENNEY, D.; OLDEKOP, J.; ASHCROFT, C.; LOCKEY, L. y PREZIOSI, R. (2005) Peer Assessment of Oral Presentations: Effects of Student Gender, University Affiliation and Participation in the Development of Assessment Criteria, *Assesment & Evaluation in Higher Education,* 30:1, pp. 21-34.

MACALPINE, J. M. K. (1999) Improving and Encouraging Peer Assessment of Student Presentations, *Assesment & Evaluation in Higher Education,* 24:1, pp. 15-25.

MACLELLAN, E. (2001) Assesment for Learning: the Differing Perceptions of Tutors and Students, *Assesment & Evaluation in Higher Education,* 26:4, pp. 307-318.

MACPHERSON, K. (1999) The Development of Critical Thinking Skills in Undergraduate Supervisory Management Units: Efficacy of Student Peer Assessment, *Assesment & Evaluation in Higher Education,* 24:3, pp. 273-284.

MAGIN, D. J. (2001a) A Novel Technique for Comparing the Reliability of Multiple Peer Assessments With That of Single Teacher Assessments of Group Process Work, *Assesment & Evaluation in Higher Education.* 26:2, pp. 139-152.

MAGIN, D. J. (2001b) Reciprocity As a Source of Bias in Multiple Peer Assessment of Group Work, *Studies in Higher Education,* 26:1, pp. 53-63.

MAGIN, D. J. y HELMORE, P. (2001) Peer and Teacher Assessments of Oral Presentation Skills; How Reliable Are They?, *Studies in Higher Education,* 26:3, pp. 287-298.

ORSMOND, P.; MERRY, S. y REILING, K. (1996) The Importance of Marking Criteria in the Use of Peer Assessment, *Assesment & Evaluation in Higher Education,* 21:3, pp. 239-250.

ORSMOND, P.; MERRY, S. y REILING, K. (2000) The Use of Student Derived Marking Criteria in Peer and Self-Assessment, *Assesment & Evaluation in Higher Education,* 25:1, pp. 23-38.

PASCUAL GÓMEZ I. y GAVIRIA SOTO J. L. (2004) El problema de la fiabilidad en la evaluación de la eficiencia docente en la universidad: una alternativa metodológica, **revista española de pedagogía,** 62:229, pp. 359-376.

REYNOLDS, M. y TREHAN. K. (2000) Assessment: a Critical Perspective, *Studies in Higher Education,* 25:3, pp. 267-278.

SEARBY, M. y EWERS, T. (1997) An Evaluation ot ihe Use of Peer Assessment in Higher Educartion: a Case Study Un the School of Music, Kingston University, *Assessment & Evaluation in Higher Education,* 22:4, pp. 371.

SULLIVAN, K. y HALL, C. (1997) Introducing Students to Self-Assessment, *Assesment & Evaluation in Higher Education.* 22:3, pp. 289-305.

TARIQ, V. N.; STEFANI, L. A. J.; BUTCHER, A. C. y HEYLINGS, D. J. A. (1998) Developing a New Approach to the Assessment of Project Work, *Assesment & Evaluation in Higher Education,* 23:3, pp. 221-240.

VVAA. (2002). SPSS 10 Guía para el análisis de datos. Servicio de informática Universidad de Cádiz.

WARD, M.; GRUPPEN, L. y REGEHR, G. (2002) Measuring Self-Assessment: Current State of the Art, *Advances in Health Sciences Education,* 7:1, pp. 63-80.

Apéndices

Apéndice A - cálculo de la fiabilidad-adaptado de Magin (2001a) y Magin y Helmore (2001).

Para el cálculo de la fiabilidad de las puntuaciones de diversos evaluadores, podemos usar los resultados del Análisis de la Varianza de un factor. Los datos que necesitamos son el número de evaluadores promedio por exposición (N) y el ratio F (suma de cuadrados dentro del grupo de evaluadores). Ambos datos son proporcionados en las tablas de resumen ANOVA de cualquier programa informático que realice estos cálculos estadísticos. Denotaremos por F* cuando se añadan, para el cálculo de los ANOVA, las puntuaciones del profesor de la asignatura a las puntuaciones de los N evaluadores.

Las variables de la investigación han sido calculadas del siguiente modo:

- Fiabilidad de las puntuaciones dentro del grupo de N evaluadores

  ❍ $r_{nn}=(F-1)/F$

- Estimación de la fiabilidad de un sólo evaluador

  ❍ $r_{ll}=(F-1)/(F+N-1)$

- Fiabilidad de las puntuaciones dentro del grupo cuando se incorporan las puntuaciones del profesor de la asignatura a los N evaluadores

  ❍ $r^*_{nn}=(F^*-1)/F^*$

- Estimación de la fiabilidad de las puntuaciones del profesor de la asignatura

  ❍ $r_{ll}=(F^*-F)/(F^*-F+1)$

- Estimación del número de evaluadores que deberían puntuar cada exposición para que su fiabilidad fuese similar a la del profesor de la asignatura

  ❍ DeltaN=N(F*-F)/(F-1)

revista española de pedagogía
año LXVII, nº 242, enero-abril 2009, 79-98

Apéndice B - plantilla para la puntuación de las presentaciones

	Niveles/puntos				Evaluación de la Presentaciones				
Criterio	**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)					

Resumen:
Los alumnos y los profesores como evaluadores. Aplicación a la calificación de presentaciones orales

La evaluación de los compañeros es una práctica que puede proporcionar interesantes ventajas, tanto motivacionales, como formativas en la enseñanza universitaria actual, que pretende ayudar a los alumnos a adquirir competencias profesionales y pretende ofrecerles cierta autonomía en su aprendizaje. El objetivo de nuestra investigación es comprobar la fiabilidad de las calificaciones de los alumnos en relación con las del profesor. Además, hemos comprobado la fiabilidad y la correlación de las calificaciones del profesor de la asignatura con las calificaciones asignadas por cuatro jueces externos. Los resultados obtenidos parecen indicar que los alumnos pueden ser unos buenos evaluadores de las presentaciones orales de sus compañeros.

Descriptores: Presentaciones orales, evaluación de compañeros, validez y fiabilidad de la evaluación.

Summary:
Students and lecturers as markers. Application in assessing oral presentations

The peer assessment is a pratice that can provide interesting advantages in the university education. Mainly if it is tried to help the students to acquire professional competencies. Our research focuses on examining the reliability of the marks given by students in relation to those given by the lecturer. In addition, we have verified the reliability and the correlation of the qualifications of the lecturer with the qualifications assigned by four external judges. The results obtained seem to indicate that students can be adequate markers for their peers' oral presentations.

Key Words: Oral presentations, Agreement among markers; Peer assessment; Reliability of the assessment.