MEDwave

Estudio primario
Medwave 2014;14(6):e5998 doi: 10.5867/medwave.2014.06.5998

Fiabilidad inter-jueces en evaluaciones portafolio de competencias profesionales en salud: el modelo de la Agencia de Calidad Sanitaria de Andalucía

Inter-rater reliability of healthcare professional skills' portfolio assessments: The Andalusian Agency for Healthcare Quality model

Autores: Antonio Almuedo-Paz[(1)], Manuel Herrera-Usagre[(1,2)], Begoña Buiza-Camacho[(1)], José Julián-Carrión[(1)], María del Pilar Carrascosa-Salmoral[(1)], Sheila María Martín-García[(1)], Rocío Salguero-Cabalgante[(1)]
Filiación:
[(1)]Agencia de Calidad Sanitaria de Andalucía, España
[(2)]Departamento de Sociología, Universidad de Sevilla, España
E-mail: manuel.herrera.usagre@juntadeandalucia.es

Citación: Almuedo-Paz A, Herrera-Usagre M, Buiza-Camacho B, Julián-Carrión J, Carrascosa-Salmoral MP, Martín-García SM, et al. Inter-rater reliability of healthcare professional skills' portfolio assessments: The Andalusian Agency for Healthcare Quality model. *Medwave* 2014;14(6):e5998 doi: 10.5867/medwave.2014.06.5998
Fecha de envío: 5/6/2014
Fecha de aceptación: 3/7/2014
Fecha de publicación: 17/7/2014
Origen: no solicitado
Tipo de revisión: con revisión por dos pares revisores externos, a doble ciego

Palabras clave: clinical competences, portfolio assessment, certification, accreditation, quality assurance

Resumen

El objetivo de este estudio es analizar la fiabilidad inter-jueces de las calificaciones realizadas por el equipo de evaluadores, pertenecientes al programa de certificación de competencias profesionales de la Agencia de Calidad Sanitaria de Andalucía (ACSA). Se estudiaron todos los procesos de certificación de competencias profesionales durante el periodo 2010-2011, independientemente de su disciplina. Se han analizado tres tipos de pruebas: 368 certificados, 17.895 informes de reflexión y 22.642 informes de práctica clínica (N = 3.010 profesionales). El porcentaje de acuerdo en las evaluaciones de certificados fue de un 89,9% (k = 0,711); 85,1% para los informes de práctica clínica (k = 0,455); y 81,7% para los informes de reflexión (k = 0,468). Los resultados de este macro-estudio muestran que la fiabilidad inter-jueces de las evaluaciones varía de ajustada a buena. En comparación con otros estudios similares, los resultados sitúan la fiabilidad del modelo en una posición cómoda. Entre las mejoras incorporadas, se incluyen la revisión de criterios y una progresiva automatización de las evaluaciones.

Abstract

This study aims to determine the reliability of assessment criteria used for a portfolio at the Andalusian Agency for Healthcare Quality (ACSA). Data: all competences certification processes, regardless of their discipline. Period: 2010-2011. Three types of tests are used: 368 certificates, 17 895 reports and 22 642 clinical practice reports (N=3 010 candidates). The tests were evaluated in pairs by the ACSA team of raters using two categories: valid and invalid. Results: The percentage agreement in assessments of certificates was 89.9%; for the reports of clinical practice, 85.1%; and for clinical practice reports, 81.7%. The inter-rater agreement coefficients (kappa) ranged from 0.468 to 0.711. Discussion: The results of this study show that the inter-rater reliability of assessments varies from fair to good. Compared with other similar studies, the results put the reliability of the

model in a comfortable position. Criteria were reviewed and progressive automation of evaluations was done.

Introducción

Un portafolio se puede definir como una colección de materiales elegidos y preparados por un profesional o estudiante, con el propósito de proveer evidencia de habilidades, conocimientos, actitudes y logros que reflejen su desempeño y actividad cotidiana [1],[2],[3]. El uso de portafolios con fines de certificación o revalidación, o como herramientas para el desarrollo profesional, está cada vez más extendido entre las distintas profesiones sanitarias [4]. El modelo de certificación de competencias profesionales de la Agencia de Calidad Sanitaria de Andalucía (ACSA) se incluye dentro de los modelos de portafolio usados para medir la competencia profesional en el ámbito sanitario [5]. Además, comparte muchos elementos con los modelos de acreditación institucional [6],[7] como la voluntariedad, la organización por estándares o la autoevaluación entre otros. Con independencia de su fin último, y como en cualquier otro instrumento de medición, se hace necesario analizar su fiabilidad y validez. Análisis que debe ser especialmente riguroso en los instrumentos orientados a la medición de la competencia clínica [8].

La fiabilidad proporciona información sobre la reproducibilidad de resultados obtenidos por un procedimiento de medición. Es el grado de estabilidad conseguido en los resultados cuando se repite una medición en condiciones idénticas. Cuando una prueba o test, además ha de ser valorada por dos o más evaluadores, se requiere verificar que existen niveles aceptables de fiabilidad en sus juicios. Se han propuesto numerosos coeficientes para cuantificar el acuerdo existente entre las medidas reportadas por dos o más observadores (jueces o evaluadores) [9]. Los procedimientos estadísticos para la evaluación del acuerdo parten de una sugerencia de Scott [10] de corregir la proporción de casos para los que el acuerdo tuvo lugar sólo por azar (*chance correction*). Varias medidas descriptivas se han definido a partir de esta sugerencia. Además del coeficiente π (Pi) introducido por Scott [10] a partir de una propuesta original de Bennett *et al*. [11], el más popular es el coeficiente κ (*Kappa*) presentado por Cohen [12],[13]. Estas medidas representan un enfoque sencillo y universalmente aceptado para medir el acuerdo, pero tienen el inconveniente de no permitir comprender la naturaleza del acuerdo y del desacuerdo. No obstante, Eye y Mun [14] se basaron en un modelo estadístico que en su aplicación práctica asumieron como válido.

Existe un considerable número de estudios que analizan la fiabilidad inter-jueces de los evaluadores sobre las pruebas aportadas por los profesionales en un modelo portafolio de certificación de competencias, existiendo una gran diversidad en el coeficiente utilizado para su análisis. Melville *et al*. [15] estudiaron la fiabilidad de un portafolio utilizado por los pediatras del *Royal College of Paediatrics and Child Health*. Para el análisis de concordancia inter-jueces utilizaron el coeficiente *Kappa* de Cohen. También utilizaron el coeficiente Kappa los estudios de Jarvis [16] y Pitts et al. [17]. Driessen *et al*. [18], en una revisión sistemática de las herramientas de portafolio utilizadas en la educación médica, excluyendo las destinadas a profesionales no médicos, concluyeron que la mayoría de los análisis de fiabilidad inter-jueces se estimaban mediante el *domain-referenced reliability coefficient* o *Kappa*, mientras que el uso del coeficiente Rho de Spearman-Brown se restringía cuando existían más de dos jueces. Gale *et al*. [19] analizaron la fiabilidad de una herramienta utilizada por un centro de selección de personal y destinada a valorar las habilidades no técnicas de los profesionales anestesistas. En su estudio analizaron la evaluación de 224 candidatos durante un periodo de dos años. Para el análisis de la concordancia inter-jueces utilizaron el coeficiente Pi de Scott, mejor que el *Kappa* de Cohen debido a que los candidatos fueron medidos por distintos evaluadores y las calificaciones de éstos se registraron de manera anónima. Otros estudios, como el de Grant *et al.* [20], se valieron del Rho para analizar un portafolio de acreditación de competencias y conocimientos para estudiantes de medicina en último curso (*Imperial College Medical School, London*). Por otro lado, Rees y Sheard [21] utilizaron el coeficiente de correlación intraclase para ver el grado de acuerdo entre dos evaluadores de los portafolios de estudiantes de la Universidad de Nottingham.

El objetivo de este estudio es analizar la fiabilidad inter-jueces de las calificaciones realizadas por el equipo de evaluadores, pertenecientes al programa de certificación de competencias profesionales de la Agencia de Calidad Sanitaria de Andalucía, sobre aquellas pruebas aportadas por los y las profesionales que desean certificar su nivel de competencia.

Métodos

Instrumento

Existen 74 modelos de portafolio de certificación de competencias profesionales, de la Agencia de Calidad Sanitaria de Andalucía. Según la evidencia científica disponible, cada uno de ellos se ajusta a los estándares de calidad sanitaria más alta en cada disciplina clínica del Sistema Sanitario Público de Andalucía (España). Cada modelo se compone de tres grandes tipos de pruebas para validar los estándares: auto-auditorías, certificados e informes.

- Auto-auditorías: son pruebas auto-evaluativas con revisiones de historias clínicas de pacientes atendidos por el o la profesional, que no necesitan presentar ningún documento de acreditación. Se realizan sobre la base de estándares establecidos por el programa [5]. Son las pruebas más comunes pero, dadas las

características del modelo de portafolio, estas pruebas no necesitan de una validación externa inicial, aunque están sujetas a posteriores auditorías.

- Certificados y otros documentos acreditativos: refrendan periodos de docencia, asistencia formativa, investigación, estancias programadas, prácticas innovadoras, entre otros, realizadas por el o la profesional en un periodo de tiempo concreto. Su validación requiere sólo del cumplimiento de una serie de requisitos explicitados claramente en los manuales de certificación.
- Informes, pueden ser de dos tipos:

1. Informes de reflexión: recogen las actuaciones concretas que él o la profesional ha realizado sobre un paciente en particular, vinculando a dicho informe su número de historia clínica, lo que permite auditar las actuaciones aportadas.
2. Informes de práctica clínica: consistentes en el análisis individual que él o la profesional hace sobre un tema concreto, por ejemplo sobre la higiene de manos o sobre los derechos de segunda generación.

Como objeto de análisis tomaremos ambos tipos de informes, así como los certificados.

Las pruebas fueron recogidas para todos los procesos de certificación de competencias profesionales, independientemente de su disciplina, durante el periodo 2010-2011 (N = 3.010 profesionales). Se han analizado un total de 368 certificados, 17.895 informes de reflexión y 22.642 informes de práctica clínica.

Estas pruebas aportadas por los y las profesionales fueron valoradas por dos evaluadores de la Agencia de Calidad Sanitaria de Andalucía de manera independiente, siguiendo una serie de criterios comunes. Los evaluadores podían calificar las pruebas utilizando dos categorías: la prueba cumple con los requisitos (prueba válida) o la prueba no cumple con los requisitos (prueba inválida). El método de evaluación de las pruebas se explica en la Tabla I.

Evaluadores					**Evaluación final**
		Iniciales			
		A	**B**	**C**	
Acuerdo	Prueba válida	+	+		+
	Prueba inválida	-	-		-
Desacuerdo	Falso positivo	+	-	-	-
	Falso negativo	+	-	+	+

Tabla I. Método de evaluación por pares de las pruebas aportadas en el modelo portafolio. *Fuente*: elaboración propia. Agencia de Calidad Sanitaria de Andalucía.

Cuando existió acuerdo entre los dos primeros evaluadores (A y B), ya sea dando la prueba de manera unánime como válida o inválida, no hizo falta la opinión de un tercer evaluador. En aquellos casos donde hubo discrepancias entre los dos primeros evaluadores se acudió a la opinión de un tercero (C), que con su juicio resolvía el estado de la prueba de manera definitiva. Así, cuando uno de los dos evaluadores (A o B) tenía una opinión minoritaria en el cómputo final, la prueba se catalogaba como "falso positivo" o bien como "falso negativo" (Tabla I), aunque esto no afectaba a la evaluación final.

Pruebas estadísticas

Se utilizarán dos estadísticos de contraste para verificar la fiabilidad inter-jueces de las dos evaluaciones: el coeficiente *Kappa* de Cohen [13], dado que es especialmente recomendable cuando se dan evaluaciones con categorías dicotómicas y dos jueces [22], como es nuestro caso. Adicionalmente, se aportará el coeficiente *Pi* de Scott o *Random Marginal Agreement Coefficient* (RMAC) [10]. Hemos añadido éste último estadístico por varios motivos. Por un lado, las características del objeto y método de evaluación son parecidas a los casos de estudio de Rees y Sheard [21] y Gale *et al* [19].

Por otro lado Krippendorff aconseja también usar el *Pi* de Scott cuando las calificaciones de los evaluadores están anonimizadas, los tamaños muestrales son grandes y los datos nominales [23], como es el caso de los dos tipos de informes. Para la aplicación de los análisis se utilizaron los software SPSS v.13 y el paquete irr del software estadístico R Package [24].

Resultados

En primer lugar, se presentan los resultados descriptivos de la concordancia entre los evaluadores por cada uno de los tipos de pruebas, a través de los porcentajes de acuerdo y desacuerdo (Tabla II). Estos porcentajes nos orientarán sobre dos aspectos, en primer lugar los datos de concordancia, es decir la proporción de acuerdo entre los dos evaluadores, ya sea para calificar la prueba como

válida o como inválida. En segundo lugar, también se presenta la proporción de casos que, tras darse el desacuerdo, un tercer evaluador desempata la calificación como "falso positivo" o como "falso negativo".

			Certificados	Informes de reflexión	Informes de práctica clínica
Concordancias entre los jueces	**Acuerdo**	Total (%)	331 (89,9)	14.621 (81,7)	19.282 (85,1)
		Pruebas Válidas (%)	267 (72,5)	12.239 (68,4)	17.251 (76,1)
		Pruebas Inválidas (%)	64 (17,4)	2.382 (13,3)	2.031 (9)
	Desacuerdo	Falsos positivos (%)	18 (4,9)	1.267 (7,1)	1.246 (5,5)
		Falsos negativos (%)	19 (5,2)	2.007 (11,2)	2.119 (9,4)
Total de pruebas analizadas (%)			368 (100)	17.895 (100)	22.642 (100)

Tabla II. Concordancia entre evaluadores. *Fuente*: ME_jora P (25) y elaboración propia.

Como se observa en la Tabla II, el porcentaje de acuerdo en las evaluaciones de certificados es superior (89,9%) al de los informes de práctica clínica (85,1%) y al de los informes de reflexión (81,7%). Los falsos negativos son la principal causa de desacuerdo en las evaluaciones entre los dos tipos de informes, mientras que no hay diferencias entre el tipo de desacuerdo que sucede en el caso de los certificados.

Tras ver los resultados descriptivos de concordancia, se vuelve necesario abordar el grado de fiabilidad inter-jueces a través de los estadísticos anteriormente descritos (Tabla II). El coeficiente *Kappa* se distribuye e interpreta como el *Alpha* de Cronbach. Es decir, tomará valores de 0 cuando se den correlaciones similares a las que se darían por azar y tomará el valor 1 cuando la correlación entre las valoraciones de ambos jueces sea perfecta. A tenor de la Tabla III, los certificados son el tipo de prueba que presenta el *Kappa* más alto (0,711) expresando una fuerza de correlación buena, según los umbrales establecidos por Sánchez-Fernández *et al.* [22]. En cambio, los coeficientes *Kappa* para los informes de reflexión y para los informes de práctica clínica son algo menores (0,468 y 0,455 respectivamente), expresando una fuerza de correlación moderada o justa.

	Certificados	Informes de reflexión	Informes de práctica clínica
% de acuerdo inter-jueces	89,9	81,7	85,1
Kappa de Cohen	0,711**	0,468**	0,455**
Pi de Scott	0,71	0,467	0,455

** p – value < 0,01.

Tabla III. Resultados de los análisis de concordancia y fiabilidad inter-jueces para los tipos de pruebas certificados, informes de reflexión e informes de práctica clínica. *Fuente*: ME_jora P (25) y elaboración propia.

El estadístico *Pi* de Scott refuerza los resultados anteriores. Siguiendo los umbrales establecidos por Landis y Koch [26], los resultados para las evaluaciones de los certificados presentan una fuerza de correlación sustancial, mientras que los coeficientes para los informes de reflexión y de práctica clínica presentan niveles moderados.

Discusión

El presente estudio es el primero que aborda la concordancia inter-jueces de las evaluaciones a doble ciego que lleva la Agencia de Calidad Sanitaria de Andalucía, para los tipos de pruebas presentadas por aquellos y aquellas profesionales que quieren certificar sus competencias. En comparación con otros estudios similares, este trabajo podría considerarse un macro-

estudio, en tanto que las muestras utilizadas son muy superiores a la media [27], como sucede con los certificados, e incluso se han llegado a utilizar varios miles de unidades muestrales, tal como ocurre con los dos tipos de informes.

Los índices *Kappa* de concordancia entre los juicios de los evaluadores que se han obtenido en el presente estudio, varían entre el considerado como bueno para los certificados (0,711) y el considerado como moderado para los informes de reflexión (0,468) y los informes de práctica clínica (0,455). Estas diferencias pueden deberse a la propia naturaleza de las pruebas analizadas. Mientras que los certificados son documentos acreditativos de una competencia, capacidad o conocimiento, por lo general requerida de manera muy precisa en la redacción del estándar –como la asistencia a un curso de formación continuada o el certificado de docencia en alguna actividad formativa-, los dos tipos de informes dan pie a una mayor subjetividad en el juicio sobre su validez, dado que son fruto del ejercicio individual y reflexivo por parte del o la profesional.

No obstante, los resultados del estudio han evidenciado índices de concordancia inter-jueces superiores a los de varios estudios similares, que además utilizan muestras mucho más pequeñas. Melville *et al.* [15] analizó la correlación inter-jueces entre dos evaluadores de modelos de portafolios de competencias para médicos pediátricos, obteniendo un *Kappa* de 0,35 para 30 casos. Por otro lado, Jarvis *et al.* [16] alcanzó un Kappa de 0,31 para la concordancia inter-jueces de 22 evaluaciones del programa de acreditación de competencias médicas generales de la *Accreditation Council for Graduate Medical Education* (ACGME). O'Sullivan *et al.* [4] midieron el grado de correlación entre los resultados de un portafolio para médicos residentes y las puntuaciones tanto clínicas como de pruebas examinadoras estandarizadas a 22 profesionales. Los autores concluyeron que la correlación era modesta (*Rho* = 0,23).

Son pocos los estudios que alcanzaron niveles de Kappa ligeramente superiores al obtenido en el presente estudio, por encima del 0,50, aun siendo inferiores al nivel considerado como bueno (0,61 - 0,80). Es el caso de los estudios de Pitts *et al.* [28], que evaluó 12 casos de un modelo de portafolio para médicos de atención primaria (κ = 0,50). También lo es el de Rees y Sheard [29], que evaluó la concordancia inter-jueces de las evaluaciones de 100 modelos de portafolios de competencias para estudiantes de medicina de segundo curso, fundado en cinco criterios. Los niveles *Kappa* variaron entre un 0,359 y un 0,693 para los distintos criterios. De los estudios contenidos en las revisiones sistemáticas sobre la validez y efectividad de los modelos de portafolios existentes hasta la fecha [30],[31], sólo el publicado por Tate *et al.* [32] obtuvo un Kappa sustancialmente superior (0,78) al obtenido en el presente estudio. Para su análisis utilizaron 100 casos y alcanzaron un porcentaje de acuerdo entre los evaluadores del 89%. Por otro lado, Grant *et al.* [20] midieron la correlación en las evaluaciones de modelos de portafolios para diferentes perfiles profesionales, alcanzando un Rho de Spearman de 0,65 como umbral más alto.

La concordancia de estas evaluaciones ha encontrado siempre grandes dificultades para mejorar las puntuaciones existentes [33]. Incluso cuando se han llevado a cabo intentos por unificar los criterios calificadores a través de cursos de formación para los evaluadores, los índices de concordancia no mejoraron [34].

Para conciliar la difícil tarea de reducir la variabilidad en las calificaciones, la mayoría de los estudios en la materia aconsejan la creación de pequeños grupos de asesores bien entrenados para el proceso evaluador; el contacto entre los jueces antes, durante y después del proceso; la elaboración de guías claras para la construcción del portafolio o una medición holística en la evaluación de los modelos de portafolios [4],[15],[29],[31],[32],[35].

La Agencia de Calidad Sanitaria de Andalucía ha mostrado una especial preocupación por la fiabilidad de las evaluaciones de los diferentes modelos de portafolio de certificación de competencias profesionales. En este sentido, la Agencia de Calidad Sanitaria de Andalucía ha implementado un progresivo proceso de automatización electrónica en las evaluaciones de los estándares, proceso que abarca en la actualidad a más del 30% de las pruebas. Del mismo modo, se han revisado los criterios de evaluación facilitando su comprensión por los evaluadores.

Entre las limitaciones del estudio destacamos dos. Primero, el estudio no ha permitido evaluar independientemente a cada evaluador, lo que resultaría de gran ayuda para focalizar mejor las intervenciones formativas relativas a la homogeneización de criterios. Segundo, futuros estudios deberán seguir realizándose con el fin de evaluar la efectividad de las medidas puestas en marcha.

Notas

Financiamiento
El presente trabajo se ha realizado con financiamiento del proyecto PI-0226 de la Convocatoria de 2010 para Proyectos de Investigación de la Conserjería de Salud y Bienestar Social de la Junta de Andalucía.

Declaración de conflictos de intereses
Los autores declaran no tener ningún conflicto de intereses en relación al trabajo presentado.

Referencias

1. Byrne M, Delarose T, King CA, Leske J, Sapnas KG, Schroeter K. Continued professional competence and portfolios. J Trauma Nurs. 2007;14(1):24-31. | PubMed |
2. Miller PA, Tuekam R. The feasibility and acceptability of using a portfolio to assess professional competence. Physiother Can. 2011;63(1):78-85. | CrossRef | PubMed |
3. Mathers NJ, Challis MC, Howe AC, Field NJ. Portfolios in continuing medical education--effective and efficient? Med Educ. 1999;33(7):521-30. | CrossRef | PubMed |
4. O'Sullivan PS, Reckase MD, McClain T, Savidge MA, Clardy JA. Demonstration of portfolios to assess competency of residents. Adv Health Sci Educ Theory Pract. 2004;9(4):309-23. | PubMed |
5. Almuedo-Paz A, Nuñez-García D, Reyes-Alcázar V, Torres-Olivera A. The ACSA Accreditation Model: self-assessment as a quality improvement tool. En: Quality Assurance. Rijeka, Croatia: Intech - Open Access Publisher, 012. | Link |
6. Pepper S. Accreditation models. Medwave. 2011 Abr;11(04):e4971 | CrossRef |
7. Pepper S. Principles of accreditation. Medwave. 2011 Mar;11(03):e4930 | CrossRef |
8. Benett Y. The Validity and reliability of assessments and self-assessments of work-based learning. Assessment & Evaluation in Higher Education. 1993;18(2):83-94. | CrossRef |
9. Ato M, Benavente A, López JJ. Análisis comparativo de tres enfoques para evaluar el acuerdo entre observadores. Psicothema. 2006;18(3):638–45. | Link |
10. Scott WA. Reliability of content analysis: the case of nominal scale coding. Public Opin Q. 1955;19(3):321. | CrossRef |
11. Bennett EM, Alpert R, Goldstein AC. Communications through limited response questioning. Public Opin Q. 1954;18(3):303-308. | CrossRef |
12. Cohen J. A coefficient of agreement for nominal scales. Ed and Psychol Meas. 1960;20(1):37–46. | CrossRef |
13. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bul. 1968;70(4):213–20.
14. Eye A von, Mun EY. Analyzing rater agreement: manifest variable methods. Mahwah, N.J.: Lawrence Erlbaum Associates, 2005.
15. Melville C, Rees M, Brookfield D, Anderson J. Portfolios for assessment of paediatric specialist registrars. Med Educ. 2004;38(10):1117-25. | CrossRef | PubMed |
16. Jarvis RM, O'Sullivan PS, McClain T, Clardy JA. Can one portfolio measure the six ACGME general competencies? Acad Psychiatry. 2004;28(3):190-6. | PubMed |
17. Pitts J, Coles C, Thomas P, Smith F. Enhancing reliability in portfolio assessment: discussions between assessors. Med Teach. 2002;24(2):197-201. | CrossRef | PubMed |
18. Driessen E, van Tartwijk J, van der Vleuten C, Wass V. Portfolios in medical education: why do they meet with mixed success? A systematic review. Med Educ. 2007;41(12):1224-33. | CrossRef | PubMed |
19. Gale TC, Roberts MJ, Sice PJ, Langton JA, Patterson FC, Carr AS, et al. Predictive validity of a selection centre testing non-technical skills for recruitment to training in anaesthesia. Br J Anaesth. 2010;105(5):603-9. | CrossRef | PubMed |
20. Grant AJ, Vermunt JD, Kinnersley P, Houston H. Exploring students' perceptions on the use of significant event analysis, as part of a portfolio assessment process in general practice, as a tool for learning how to use reflection in learning. BMC Med Educ. 2007;7(1):5. | CrossRef | PubMed |
21. Rees CE, Sheard CE. The reliability of assessment criteria for undergraduate medical students' communication skills portfolios: the Nottingham experience. Med Educ. 2004;38(2):138-44. | CrossRef | PubMed |
22. Sánchez Fernández P, Aguilar de Armas I, Fuentelsaz Gallego C, Teresa Moreno Casbas M, Hidalgo García R. Fiabilidad de los instrumentos de medición en ciencias de la salud. Enf Clín. 2005;15(4):227–36. | CrossRef |
23. Krippendorff K. Reliability in content analysis: some common misconceptions and recommendations. Human Communication Research. 2004;30(3):411–33. | CrossRef |
24. Gamer M, Lemon J, Fellows I, Puspendra S. irr: various coefficients of interrater reliability and agreement. 2012. cran.r-project.org [on line] | Link |
25. Agencia de Calidad Sanitaria de Andalucía. ME_jora P [Internet]. Accreditation of Professional Competences (in spanish). 2012 [on line]. | Link |
26. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-74. | PubMed |
27. Driessen EW, Muijtjens AM, van Tartwijk J, van der Vleuten CP. Web- or paper-based portfolios: is there a difference? Med Educ. 2007;41(11):1067-73. | CrossRef | PubMed |
28. Pitts J, Coles C, Thomas P. Educational portfolios in the assessment of general practice trainers: reliability of assessors. Med Educ. 1999;33(7):515-20. | CrossRef | PubMed |
29. Rees CE, Sheard CE. The reliability of assessment criteria for undergraduate medical students' communication skills portfolios: the Nottingham experience. Med Educ. 2004 Feb;38(2):138-44. | CrossRef | PubMed |
30. Driessen E, van Tartwijk J, van der Vleuten C, Wass V. Portfolios in medical education: why do they meet with mixed success? A systematic review. Med Educ. 2007;41(12):1224-33. | CrossRef | PubMed |
31. Overeem K, Faber MJ, Arah OA, Elwyn G, Lombarts KM, Wollersheim HC, et al. Doctor performance assessment in daily practise: does it help doctors or not? A systematic review. Med Educ. 2007;41(11):1039-49. | CrossRef | PubMed |
32. Tate P, Foulkes J, Neighbour R, Campion P, Field S. Assessing physicians' interpersonal skills via videotaped encounters: a new approach for the Royal College of General Practitioners Membership

examination. J Health Commun. 1999;4(2):143-52. | CrossRef | PubMed |
33. Pitts J, Coles C, Thomas P. Educational portfolios in the assessment of general practice trainers: reliability of assessors. Med Educ. 1999;33(7):515-20. | CrossRef | PubMed |
34. Pitts J, Coles C, Thomas P. Enhancing reliability in portfolio assessment: 'shaping' the portfolio. Med Teach. 2001;23(4):351-356. | CrossRef | PubMed |
35. McCready T. Portfolios and the assessment of competence in nursing: a literature review. Int J Nurs Stud. 2007;44(1):143-51. | CrossRef | PubMed |

Correspondencia a:
C/ Augusto Peyré
Nº 1 Edificio Olalla, 3ª Planta. 41020
Sevilla
España