Limitaciones Metodológicas y Soluciones Factibles en la Valoración y Cálculo de las Calificaciones obtenidas Mediante las Rúbricas como Estrategias de Evaluación de Competencias de los Estudiantes Universitarios y no Universitarios

- Clemente Rodriguez-Sabiote [*]
- José Alvarez-Rodríguez [**]
- Rosa del Pilar Gámez-Durán [***]

Resumen

El diseño y la elaboración de rúbricas como estrategias de evaluación en enseñanza superior y otros niveles educativos, así como los numerosos beneficios de su utilización son temáticas ampliamente tratadas en la literatura científica. Sin embargo, los aspectos relativos al cálculo de las puntuaciones que se derivan de su uso, así como la fiabilidad y validez de las rúbricas han tenido una menor atención. El presente trabajo, por consiguiente, se centra en la exposición de un conjunto de limitaciones metodológicas relacionadas con las rúbricas y en la propuesta de soluciones factibles para superarlas. Las consideraciones finales del trabajo orientan propuestas relativas a rúbricas donde las puntuaciones no sean equivalentes al sistema de calificación del país de referencia, ni a su ponderación. Se destacan, también, los sesgos derivados de la utilización de coeficientes de correlación y fiabilidad en el uso de la fiabilidad intra e inter-observadores y en la validez de contenido y conveniencia de su sustitución por coeficientes de concordancia.

Palabras clave: Rúbricas. Evaluación de competencias. Fiabilidad y validez.

[*] Doctor en Pedagogía por la Universidad de Granada.Profesor Titular de Universidad del departamento de Métodos de Investigación y Diagnóstico de la Universidad de Granada (España; E mail; clerosa@ugr.es.

[**] Doctor en Pedagogía por la Universidad de Granada. Profesor Titular de Univesidad del departamento de Pedagogía de la Universidad de Granada (España); E mail: alvarez@ugr.es.

[***] Diplomada en Educación Primaria por la Universidad de Granada. Maestra de Primaria con plaza de funcionaria en la Junta de Andalucía (España); E mail: rosanagamez72@gmai.com.

1. Introducción

No son pocos los trabajos donde se desarrollan interesantes disertaciones teóricas y empíricas sobre los beneficios de las rúbricas como instrumentos de evaluación de competencias de estudiantes no universitarios y universitarios. Sirvan como ejemplo los trabajos más actuales de ArchMiller et al. (2017), Dickinson& Adams (2017), El Turkey et al. (2017), Kettler&Bower (2017), Martínez-Figueira, Tellado González y Raposo Rivas (2013), Murillo-Zamorano &Montanero (2017), Smit et al. (2017), Wessel, McDonald &Cebrian (2017), Winder et al. (2017); Korkut (2017) y Yamamoto, Umemura& Kawano (2017). En todos ellos se tratan aspectos de las rúbricas relacionados con el diseño, la planificación y elaboración de las mismas, así como de las diferencias de resultados obtenidos respecto a otros instrumentos de evaluación de competencias de los niveles universitarios y no universitarios.

Además, en muchos de estos trabajos, como los de Valverde-Berrocoso y Ciudad-Gómez (2014), Ibarra & Rodríguez y Gómez, (2012) y Rodríguez & Ibarra (2017) se concluye que las rúbricas son, efectivamente, instrumentos de evaluación para el aprendiza¬je, sostenibles, orientados a la valoración de las competencias y con un papel activo del estudiante y que, también, pueden aportar un gran potencial para la autorregulación y autoeficacia del mismo (del estudiante), para la calidad de la evaluación y mejora de la docencia.

Sin embargo, son más escasos los trabajos donde se tratan aspectos relacionados con la fiabilidad de las medidas de los evaluadores y la validez del contenido de los aspectos o rubros de medición de la rúbrica. Destacamos, no obstante, entre los trabajos más recientes, los de Henderson, (2016), Merma Molina, Peña Alfaro & Peña (2017), Park et al. (2016), Roegman et al. (2016), Valverde-Berrocoso y Ciudad-Gómez (2014); Wesolowski et al. (2017).

Por consiguiente, no es extraño que la despreocupación por la validez y la fiabilidad sea la principal crítica que encontraron Reddy & Andrade (2010) al concluir los estudios evaluados por ellos y notar que no se hacía mención al proceso de desarrollo de las rúbricas para establecer su calidad (CANO, 2015, p. 273).

Así pues, dejando a un lado los complejos procedimientos relacionados con el diseño y elaboración de las rúbricas, aspecto no relacionado con el presente trabajo, nos vamos a centrar en la problemática metodológica concerniente al cálculo de las puntuaciones y la fiabilidad de los juicios emitidos por los evaluadores y la validez de los aspectos o rubros medidos por la rúbrica. En este sentido, destacamos cuatro problemas principales y sus posibles soluciones.

2. Limitaciones metodológicas en la valoración y cálculo de las puntuaciones de las rúbricas

2.1. El problema número 1: la transformación de los valores de logro al sistema de calificación de referenciao

El primer obstáculo que nos encontramos tras aplicar y valorar mediante puntuaciones una rúbrica es el de transformar los valores de logro de la misma (1 a 3;1 a 5;1 a 7, etc.) al sistema de calificación de referencia del país donde se realice nuestra evaluación, sólo en el caso, evidentemente, que dichos valores no coincidan con el sistema de calificación contemplado. Como sabemos, no todos los países tienen el mismo sistema de calificación y ni siquiera en muchos casos este es de tipo numérico exclusivo (véase el caso de Estados Unidos de América y otros países afines).

2.2. El problema número 2: la ponderación de cada uno de los niveles de logro en porcentajes o proporciones correspondientes en relación con la calificación final (sólo para rúbricas analíticas)

Como hemos señalado en un apartado anterior existen dos tipologías de rúbricas fundamentales, a saber, globales y analíticas. En las primeras (globales u holísticas) se desarrolla una valoración integrada sin determinar componentes del proceso o tema evaluado. En las segundas (analíticas) se implementa, por el contrario, una valoración de partes del desempeño del alumnado desglosando sus componentes, indicadores o rubros. Precisamente este último hecho (el desglose en dichos rubros) conforma un segundo problema metodológico, dado que habrá que ponderar cada uno de los niveles de logro, bien en porcentajes, bien en proporciones que le corresponden a cada rubro en relación con la calificación final.

2.3. El problema número 3: la concordancia intra-evaluadores (para un solo evaluador) y entre-evaluadores en el caso de que haya más de un evaluador

Hablar de limitaciones de fiabilidad de las valoraciones realizadas por los diferentes evaluadores no es algo que pueda considerarse novedoso. Interesantes aportaciones, a este respecto, ya fueron destacadas por Black (1998), Davidson, Howel & Hoekema (2000), Moskal & Leydens (2000) y Shavelson, Gao & Baxter (1996). No obstante, en niveles educativos no universitarios no es habitual que una rúbrica sea valorada por más de un evaluador, razón por la cual la calificación exclusiva corresponde a un solo evaluador sin que sea posible, por tanto, la problemática de la concordancia entre-evaluadores, aunque sí la del propio evaluador consigo mismo (intra-evaluador). Sin embargo, en niveles educativos universitarios el uso de rúbricas como instrumento de valoración de trabajos de fin de grado, fin de máster e incluso tesis de doctorado, donde concurren las valoraciones individuales de varios evaluadores, puede plantear un importante problema de sesgo cuando las valoraciones son evidentemente distintas. Algunos pueden pensar que la solución a tal incidencia puede pasar por el cálculo promedio de las valoraciones de los distintos evaluadores o del promedio de puntuaciones de un único evaluador. Tal procedimiento es el adecuado y conveniente, pero sólo cuando, previamente, se ha contrastado una concordancia entre-evaluadores / intra-observadores lo suficientemente importante, como para no deberse al mero acuerdo por azar. A este respecto, en un excelente trabajo de Cano (2015, p. 272) se muestra una síntesis completa y convenientemente documentada sobre posibles estrategias de abordaje de la fiabilidad de las rúbricas a partir de un espléndido trabajo de Jonsson & Svingby (2007, p. 134). En este último trabajo los autores describen toda una gama de estrategias de garantía de la fiabilidad entre e intra-observadores que posteriormente abordaremos con mayor detalle.

2.4. El problema número 4: la validez de la rúbrica elaborada

El cuarto problema se relaciona con la validez de la rúbrica, o dicho de otra forma, en qué grado la rúbrica mide las competencias, los objetivos de aprendizaje, etc. para lo cual ha sido elaborada. Dicho concepto se refiere a lo que desde la Teoría Clásica de los Tests (TCT) se denominan las diferentes tipologías de validez, es decir, la validez de contenido,

de constructo y criterial concurrente y predictiva. Por otra parte, en los trabajos de Moni, Beswick & Moni, 2005; Green & Browser, 2006 y Lapsley & Moody, 2007, citados por Cano (2015, p. 273), se pone de manifiesto que, como ocurría con la fiabilidad, la validez es un aspecto desafortunadamente descuidado o tratado superficialmente. No obstante, Messick (1996) cita incluso seis tipos diferentes aspectos de validez: contenido, generalizabilidad, externalidad, estructuralidad, sustantividad y consecuencial que han sido contemplados sobre las rúbricas hasta un total de 25 estudios, según muestran sendos trabajos Jonsson & Svingby (2007, p. 136) y Cano (2015, p. 272).

3. Soluciones posibles para la transformación de los valores de logro al sistema de calificación de referencia y la ponderación de cada uno de los niveles de logro en relación con la calificación final (sólo para rúbricas analíticas)

Para los problemas 1 y 2, es decir, para la transformación de los valores de logro al sistema de calificación de referencia y la ponderación de cada uno de los niveles de logro en porcentajes o proporciones que le correspondan en relación con la calificación final (solo para rúbricas analíticas), proponemos el uso de distintas ecuaciones y su correspondientes algoritmos de cálculo en un recurso como Excel para, lógicamente, ganar tiempo y precisión en el cálculo de la calificación final del estudiante. Sin embargo, antes de mostrar estas ecuaciones, proponemos la siguiente tabla donde se contemplan los posibles casos que pueden darse, según la correspondencia tridimensional de tres elementos, a saber: número de evaluadores (1 o 2 o más); niveles de logro de la rúbrica (no equivalentes o equivalentes a puntuaciones estándar) y el tipo de rúbrica contemplada (global o analítica). Dicha correspondencia abarcaría un total de 8 casos diferentes que serían los siguientes:

Tabla 1- Diferentes tipos de casos de rúbrica según la correspondencia de las variables número de evaluador-es, niveles de logro equivalentes o no equivalente a puntuaciones estándar, así como el tipo de rúbrica contemplado

Número de evaluadores	**Niveles de logro no equivalentes a puntuaciones estándar**		**Niveles de logro equivalentes a puntuaciones estándar**	
	Rúbrica global	**Rúbrica analítica**	**Rúbrica global**	**Rúbrica analítica**
1 evaluador	Caso tipo I	Caso tipo II	Caso tipo V	Caso tipo VI
≥2 evaluadores	Caso tipo III	Caso tipo IV	Caso tipo VII	Caso tipo VIII

Fuente: Elaboración propia (2018).

Para cada uno de estos 8 casos proponemos una ecuación de cálculo de la puntuación final del estudiante.

3.1. Rúbricas con niveles de logro no equivalentes a calificaciones estándar[1]

Rúbricas globales

A) Para un evaluador (caso tipo I)

Ecuación para el cálculo de calificación de un estudiante en rúbricas con niveles de logro no equivalentes a calificaciones estándar para el caso de las rúbricas globales y un evaluador

$$Calif. = Punt.evaluador\ único \times \left(\frac{10}{nºmáx.nivel\ logro\ rúbrica}\right) \quad (1)$$

B) Para dos o más evaluadores (caso tipo III)

Ecuación para el cálculo de calificación de un estudiante en rúbricas con niveles de logro no equivalentes a calificaciones estándar para el caso de las rúbricas globales y dos o más evaluadores

$$Calif. = \sum_{i=1}^{n}\left(\frac{Pev1+Pev2+Pev3+\cdots pevn}{nºevaluadores}\right) \times \left(\frac{10}{nºmáx.nivel\ logro\ rúbrica}\right) \quad (2)$$

Rúbricas analíticas o ponderadas

C) Para un evaluador (caso tipo II)

Ecuación para el cálculo de calificación de un estudiante en rúbricas con niveles de logro no equivalentes a calificaciones estándar para el caso de las rúbricas analíticas y un evaluador.

$$Calif. = \sum_{i=1}^{n} Punt.eval.único \times \left(\frac{Punt.nivel\ logro\ en\ calif.estándar}{nºniveles\ de\ logro\ rúbrica}\right) \quad (3)$$

D) Para dos o más evaluadores (caso tipo IV)

[1] En el ejemplo se utiliza el sistema de calificación español que va desde 0 a 10, razón por la cual el numerador del segundo miembro de la ecuación es 10 (puntuación máxima). Obviamente puede cambiarse para el país donde pretenda utilizarse sustituyéndolo por el valor máximo del sistema de calificación de dicho país de referencia.

Ecuación para el cálculo de la calificación de un estudiante en rúbricas con niveles de logro no equivalentes a calificaciones estándar para el caso de las rúbricas analíticas y dos o más evaluadores

$$Calif. = \sum_{i=1}^{n} \left(\frac{Pev1+Pev2+Pev3+\cdots pevn}{nºevaluadores}\right) \times \left(\frac{Punt.nivel\ logro\ en\ calf.estándar}{nºniveles\ de\ logro\ rúbrica}\right) \quad (4)$$

3.2. Rúbricas con niveles de logro equivalentes a calificaciones estándar

Rúbricas globales

A) Para un evaluador (caso tipo V)

Ecuación para el cálculo de la calificación de un estudiante en rúbricas con niveles de logro equivalentes a calificaciones estándar para el caso de las rúbricas analíticas y un solo evaluador

$$Calif. = Punt.evaluadorúnico \quad (5)$$

B) Para dos o más evaluadores (caso tipo VII)

Ecuación para el cálculo de la calificación de un estudiante en rúbricas con niveles de logro equivalentes a calificaciones estándar para el caso de las rúbricas globales y dos o más evaluadores.

$$Calif. = \sum_{i=1}^{n} \left(\frac{Pev1+Pev2+Pev3+\cdots pevn}{nºevaluadores}\right) \quad (6)$$

Rúbricas analíticas o ponderadas

C) Para un evaluador (caso tipo VI)

Ecuación para el cálculo de la calificación de un estudiante en rúbricas con niveles de logro equivalentes a calificaciones estándar para el caso de las rúbricas analíticas y un solo evaluador.

$$Calif. = \sum_{i=1}^{n} (Punt.ev.único\ x\ Punt.ponderada\ en\ proporción) \quad (7)$$

D) Para dos o más evaluadores (caso tipo VIII)

Ecuación para el cálculo de la calificación de un estudiante en rúbricas con niveles de logro equivalentes a calificaciones estándar para el caso de las rúbricas analíticas y dos o más evaluadores.

$$Calif. = \sum_{i=1}^{n} \left(\frac{Pev1+Pev2+Pev3+\cdots pevn}{nºevaluadores}\right) \times (Punt. ponderada\ en\ proporción) \quad (8)$$

4. Soluciones viables para el problema de la concordancia entre-evaluadores compatibles con la fiabilidad y la validez de contenido conjuntamente

Como se ha reseñado en un apartado anterior el problema de la concordancia entre evaluadores no es improbable, habida cuenta de la opinión particular y subjetiva con que algunos niveles de logro pueden valorarse por cada uno de los evaluadores individualmente. Por consiguiente, proponemos como criterio previo a cualquier cálculo promedio entre-evaluadores la comprobación del grado de concordancia entre los mismos teniendo en cuenta como premisa básica de partida que estamos tratando con variables numéricas continuas y otras veces con variables de naturaleza categórica o nominal. Y es importante reseñar el término concordancia, porque como destacan Jonsson & cSvingby (2007, p. 134) citados, y también antes mencionado sobre el trabajo de Cano (2015, p. 272), la fiabilidad de las puntuaciones de una rúbrica ha sido abordada desde diferentes prismas. Los principales han sido la consistencia: coeficientes de correlación y fiabilidad, etc.; la estimación de las medidas: Teoría de la Generalizabilidad (TG), Teoría de Respuesta al Ítem (TRI), etc., así como desde la concordancia. Todos ellos tendrían relación, además, con el parámetro de validez de contenido propuesto desde la Teoría Clásica de los Tests, razón por la cual los abordamos conjuntamente. No son pocos los trabajos, entre otros los de Brown, Glaswell y Harland (2004), Marzano (2000), Stemler (2004), Stoddart et al. (2000), donde se explicitan los valores mínimos aceptables para la consecución de la fiabilidad de las valoraciones efectuadas.

Sin embargo, en el presente trabajo defendemos la eliminación de los coeficientes de correlación y fiabilidad porque pueden crear ilusión de acuerdo sin que este sea real (CORTÉS-REYES; RUBIO-ROMERO; GAITÁN-DUARTE, 2010; ESQUIVEL-MOLINA et al., 2006, PEDROSA; SUÁREZ-ÁLVAREZ; GARCÍA-CUETO, 2013y MARTÍNEZ-CURBELO; CORTÉS-CORTÉS; PÉREZ-FERNÁNDEZ, 2016), de dispersión porque su interpretación es subjetiva (ABAD et al., 2011; CLAEYS et al.,2012) y, por consiguiente, proponemos como

estimadores de la fiabilidad y validez de contenido de las rúbricas a los coeficientes de concordancia. Con mayor detalle recogemos en la siguiente tabla los que, a nuestro juicio, son los coeficientes de concordancia de mayor utilidad para este fin:

Tabla 2- Propuesta de coeficientes de concordancia de mayor utilidad para el cálculo de la fiabilidad y la validez de contenido

Coeficientes	**Escala de medida**	**Número de evaluadores**
Kappa de Cohen	Nominal y ordinal	2
Kappa de Fleiss	Nominal y ordinal	≥2
Coeficiente de Correlación Intraclase (CCI)	Intervalo (continua)	≥2
Bland-Altman	Intervalo (continua)	2

Fuente: Elaboración propia (2018).

La premisa básica de ello reside en que correlación no es concordancia, aunque en apariencia se parezcan y, a veces, correlación y fiabilidad con respecto a concordancia coincidan (situación I o verdadero acuerdo) o no (situación II o falso acuerdo).

Tabla 3- Posibles situaciones cuando convergen, o no, la correlación y fiabilidad con la concordancia

Posibles situaciones	**Correlación y fiabilidad**	**Concordancia**
Correlación y fiabilidad		**Situación I** Corr. & Fiab. convergen con Concord. **Verdadero acuerdo**
Concordancia	**Situación II** Corr. & Fiab. divergen con Concord. **Falso acuerdo**	

Fuente: Elaboración propia (2018).

Así, mientras la correlación es covariación entre series de números que proponen dos evaluadores, es decir, jerarquías ordinales en su posicionamiento y la fiabilidad, el grado de consistencia de las puntuaciones, por el contrario, la concordancia significa que las valoraciones de uno y otro evaluador deben ser similares y que se logrará más o menos concordancia cuanto mayor o menor se parezcan dichas valoraciones. Veamos dos ejemplos sencillos para su comprensión. Imaginemos las puntuaciones de dos

evaluadores a diez y cinco estudiantes en una rúbrica de tres y diez niveles de logro respectivamente (situaciones tipo I y II)[2].

Tabla 4- Valoraciones de dos evaluadores a diez estudiantes donde correlación y concordancia convergen (situación tipo I).

Estudiante	Evaluador 1	Evaluador 2
A	1	2
B	2	2
C	1	1
D	3	2
E	1	1
F	1	1
G	2	2
H	3	3
I	1	1
J	2	2

Fuente: Elaboración propia (2018).

En la situación de convergencia (situación tipo I) los valores de correlación (Pearson, Kendall y Spearman) estarían comprendidos entre r = 0.80 ↔ 0.82 y el valor de α = 0.89 y, en todo caso, serían estadísticamente significativos (p<.01). Por otra parte, el valor de concordancia entre los dos observadores calculado mediante el coeficiente de Kappa de Cohen ascendería a K=0.68 y estaría asociado a una puntuación típica t= 3.03 y, por tanto, estadísticamente significativa (p=0.02). En los dos casos podríamos concluir por vías diferentes y convergentes (caso tipo I) que como existe alta correlación entre los dos evaluadores y también una alta concordancia, aunque, como veremos ello no siempre es así.

Tabla 5 - Valoraciones de dos evaluadores a cinco estudiantes donde correlación y concordancia divergen (situación tipo II).

Estudiante	Evaluador 1	Evaluador 2
A	1	2
B	3	4
C	5	6
D	7	8
E	9	10

Fuente: Elaboración propia (2018).

[2] Todos los cálculos implementados se han llevado a cabo mediante el programa SPSS. v.24.

Por el contrario, en una situación de divergencia (situación tipo II) los valores de correlación (correlaciones de Pearson, Kendall y Spearman) y de fiabilidad como consistencia interna (α de Cronbach) en nuestro caso particular ascenderían a r= 1, es decir, que habría correlación perfecta, mientras el valor de alfa de Cronbach daría como resultado una fiabilidad perfecta de α =1. Por otra parte, el valor de concordancia entre los dos observadores calculado mediante el coeficiente de Kappa de Cohen ascendería a K= 0. En el primer caso, por la vía de correlación y fiabilidad como consistencia interna podríamos concluir que, dado que los coeficientes de correlación y el coeficiente de fiabilidad indican la unidad (1), las valoraciones de los evaluadores han logrado un acuerdo perfecto, cuando puede apreciarse que nos encontraríamos en una situación diametralmente opuesta. Una posible solución podría pasar por el uso del coeficiente de concordancia de Kappa de Cohen que sí que permitiría determinar la nula concordancia de los 2 evaluadores a la hora de valorar a los 5 estudiantes.

5. Consideraciones finales

La planificación, utilización y beneficios de las rúbricas como instrumentos de evaluación de competencias del alumnado de educación universitaria y otros niveles ha tenido una prolífica producción científica a lo largo de estos últimos años, o, lo que es lo mismo, ha sido un tópico de investigación de considerable importancia. Sin embargo, no puede afirmarse que los procedimientos de cálculo de las calificaciones mediante tales recursos de evaluación, ni la fiabilidad o validez cuando hay más de un evaluador midiendo a idéntico alumnado hayan obtenido similar atención. En el presente trabajo, por consiguiente, hemos implementado una propuesta con la firme convicción de que puede ser útil para dotar de mayor agilidad a los procesos de cálculo de calificaciones mediante las rúbricas, así como para garantizar la fiabilidad y validez de las rúbricas cuando hay más de un evaluador.

En este sentido, en relación con la primera cuestión (procedimiento de cálculo de las rúbricas) hemos propuesto una serie de ecuaciones para su cálculo que dependen del tipo de caso en que nos encontremos. A este respecto, hemos creado hasta ocho tipologías diferentes (caso tipo I … hasta caso tipo VIII) por correspondencia tridimensional al cruzar: número de evaluadores (1 o 2 o más) por niveles de logro equivalentes o no a las

puntuaciones del sistema de calificación del país de referencia por tipo de rúbricas (analíticas vs globales) incluyendo recursos de análisis que facilitan el procedimiento de cálculo como Excel.

En relación con la segunda cuestión, no menos importante, relativa a la fiabilidad y validez de las rúbricas cuando hay 2 o más evaluadores hemos mostrado como los procedimientos para su consecución son abordados en muchos casos mediante estrategias sesgadas que incluyen la correlación y la fiabilidad como consistencia interna. El problema de fondo que subyace en todo ello radica en confundir correlación y fiabilidad como consistencia interna con concordancia entre observadores. Dicha confusión resulta inocua en los casos que nosotros hemos denominado situaciones tipo I (convergencia), que es cuando las tres estrategias (correlación, fiabilidad como consistencia interna y concordancia coinciden o convergen en sus resultados y conclusiones que de ellos se derivan). En este tipo de situaciones el grave sesgo de la utilización de coeficientes de correlación y fiabilidad por los de concordancia queda diluido. Por el contrario, queda en evidencia en las situaciones tipo II (divergencia) cuando realmente somos capaces de delimitar las diferencias entre los tres conceptos y la inconveniencia del uso de los coeficientes de correlación y fiabilidad por la creación de lo que en el presente trabajo podemos denominar falsa ilusión de acuerdo o falso positivo.

Referências

ABAD, F. et al. *Medición en ciencias sociales y de la salud.* Madrid: Síntesis, 2011.

ARCHMILLER, A. et al. Group peer assessment for summative evaluation in a graduate-level statistics course for ecologists. *Assessment and Evaluation in Higher Education*, v. 8., n. 42, p. 1208-1220, 2017.

BLACK, P. *Testing:* Friend or foe? London: Falmer Press, 1998.

BROWN, G. T. L.; GLASSWELL, K.; HARLAND, D. Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, n. 9, p. 105–12, 2004.

CANO, E. Las rúbricas como instrumento de evaluación de competencias en Educación Superior ¿Uso o abuso? *Profesorado:* Revista de Curriculum y Formación del Profesorado, v. 19, n. 2, p. 266-280, 2015.

CLAEYS, C. et al. Content validity and inter-rater reliability of an instrument to characterize unintentional medication discrepancies. *Drugs Aging*, n. 29, p. 577-591, 2012.

CORTÉS-REYES, E.; RUBIO-ROMERO. J. A.; GAITÁN-DUARTE, H. Métodos estadísticos de evaluación dela concordancia y la reproducibilidad de pruebas diagnósticas. *Revista Colombiana de Obstetricia y Ginecología*, v. 61, n. 3, p. 247-255, 2010.

DAVIDSON, M.; HOWELL, K. W.; HOEKEMA, P. Effects of ethnicity and violent content on rubric scores in writing samples. *Journal of Educational Research*, v. 93, p. 367–373, 2000.

DICKINSON, P.; ADAMS, J. Values in evaluation: The use of rubrics. *Evaluation and Program Planning,* v. 65, p. 113-116, 2017.

EL TURKEY, H. et al. The Creativity-in-Progress Rubric on Proving: Two Teaching Implementations and Students' Reported Usage. *PRIMUS*, p. 1-23, 2017.

ESQUIVEL-MOLINA, C. et al. Coeficiente de correlación intraclase *vs* correlación de Pearson de langlucemia capilar por reflectometría y glucemia plasmática. *Medicina Interna de México*, v. 22, n. 3, p. 165-170, mayo-jun., 2006.

HENDERSON, S. J. et al. Validation of Assessment Vignettes and Scoring Rubric of Multicultural and International Competency in Faculty Teaching. *Multicultural Learning and Teaching*, v. 11, n. 1, p. 53-81, 2016.

IBARRA SÁIZ, M. S.; RODRÍGUEZ GÓMEZ, G.; GÓMEZ RUIZ, M. A. La evaluación entre iguales: beneficios y estrategias para su práctica en la universidad. *Revista de Educación*, v./n. 359, p. 206-231, set./dez., 2012.

JONSSON, A.; SVINGBY, G. The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, n. 2, p. 130–144, 2007.

KETTLER, .; BOWER, J. Measuring Creative Capacity in Gifted Students: Comparing Teacher Ratings and Student Products. *Gifted Child Quarterly* , [S.l.], v. 61, n. 4, p. 290-299, 2017.

KORKUT, P. The construction and pilot application of a scoring rubric for creative drama lesson planning. *Research in Drama Education*, p. 1-12., 2017.

MARTÍNEZ-CURBELO, G.; CORTÉS-CORTÉS, M.; PÉREZ-FERNÁNDEZ, A. Metodología para el análisis de correlación y concordancia en equipos de mediciones similares. *Universidad y Sociedad*, v. 8, n. 4., p. 65-70, 2016.

MARTÍNEZ-FIGUEIRA, M. E.; TELLADO GONZÁLEZ, F.; RAPOSO RIVAS, M. La rúbrica como instrumento para la autoevaluación: un estudio piloto. *Revista de Docencia Universitaria,* v. 11, n. 2, p. 373-390, mayo-ago. 2013.

MARZANO, R. J. A comparison of selected methods of scoring classroom assessments. *Applied Measurement in Education*, v. 15, p. 249–267, 2002.

MERMA MOLINA, G.; PEÑA ALFARO, H.; PEÑA, A.G. Design and Validation of a Rubric to Assess the Use of American Psychological Association Style in Scientific Articles. *Journal of New Approaches in Educational Research,* [S.l.], v. 6, n. 1, p. 78-86, 2017.

MESSICK, S. Validity of Performance Assessments. In: PHILLIPS, G. W. (Ed.). *Technical Issues in Large-Scale Performance Assessment*. Washington, DC: NCES -National Center for Education Statistics, 1996.

MOSKAL, B. M.; LEYDENS, J. A. Scoring rubric development: Validity and reliability. *Practical Assessment*, *Research & Evaluation*, v. 7, p. 71–81, 2000.

MURILLO-ZAMORANO, L. R.; MONTANERO, M. Oral presentations in higher education: a comparison of the impact of peer and teacher feedback Assessment and Evaluation in Higher Education, v. 43, n. 1, p. 138-150, 2017.

PARK, Y. S. et al. Inter-Rater Reliability and Generalizability of Patient Note Scores Using a Scoring Rubric Based on the USMLE Step-2 CS Format. *Advances in Health Sciences Education*, [S.l.], v. 21, n. 4, p. 761-773, 2016.

PEDROSA, I; SUÁREZ-ÁLVAREZ, J.; GARCÍA-CUETO, E. Evidencias sobre la validez de contenido: avances teóricos y métodos para su estimación. *Acción Psicológica*, v. 10, n. 2, p. 78-89, 2013.

REDDY, Y., ANDRADE, H. A review of rubric use in higher education. *Assessment & Evaluation In Higher Education*, n.35, p. 435- 448, 2010.

RODRÍGUEZ GÓMEZ, G.; IBARRA SAIZ, M. S. Reflexiones en torno a la competencia evaluadora del profesorado en la Educación Superior. *REDU: Revista de Docencia Universitaria*, v. 10, n. 2, p. 149-161, 2017.

ROEGMAN, R. et al. Unpacking the Data: An Analysis of the Use of Danielson's (2007) "Framework for Professional Practice". Teaching Residency Program. *Educational Assessment, Evaluation and Accountability*, v. 28, n. 2, p. 111-137, 2016.

SHAVELSON, R. J.; GAO, X. & BAXTER, G. On the content validity of performance assessments: Centrality of domain-specifications. In: BIRENBAUM, M.; DOCHY, F. (Ed.). *Alternatives in assessment of achievements, learning processes and prior knowledge*. Boston: Kluwer Academic Publishers, 1996.

SMIT, R. et al. Effects of a Rubric for Mathematical Reasoning on Teaching and Learning in Primary School. *Instructional Science: An International Journal of the Learning Sciences*, v. 45, n. 5, p. 603-622, 2017.

STEMLER, S. E. A comparison of consensus, consistency, and measurement approaches to estimating inter-rater reliability. *Practical Assessment, Research & Evaluation*, v. 9. n. 4, 2004.

STODDART, T. et al . Concept maps as assessment in science inquiry learning: A report of methodology. *International Journal of Science Education*, n. 22, p. 1221–1246, 2000.

VALVERDE-BERROCOSO, J.; CIUDAD-GÓMEZ, A. El uso de e-rúbricas para la evaluación de competencias en estudiantes universitarios. Estudio sobre fiabilidad del instrumento. *REDU: Revista de Docencia Universitaria*, Número monográfico dedicado a Evaluación formativa mediante E-rúbricas, v. 12, n. 1, p. 49-79, 2014.

WESOLOWSKI, B. C. et al. The Development of a Secondary-Level Solo Wind Instrument Performance Rubric Using the Multifaceted Rasch Partial Credit Measurement Model. *Journal of Research in Music Education*, v. 65, n. 1, p. 95-119, 2017.

WESSEL, C. C., MC DONALD, A.; CEBRIAN, J. An evaluative tool for rapid assessment of derelict vessel effects on coastal resources. *Journal of Environmental Management*, n. 207, p. 262-268, 2017.

WINDER, C. B. et al. Comparison of online, hands-on, and a combined approach for teaching cautery disbudding technique to dairy producers. *Journal of Dairy Science*, 2017.

YAMAMOTO, M.; UMEMURA, N.; KAWANO, H. Automated essay scoring system based on rubric. *Studies in Computational Intelligence*, n. 727, p. 177-190, 2017.

Recebido em: 16/01/2018

Aceito para publicação em: 04/11/2018

Methodological limitations and feasible solutions in the assessment and grade calculation obtained through the rubrics as skills assessment strategies of higher education students and other levels

Abstract

The design and elaboration of rubrics as assessment strategies in higher education and other educational levels, as well as the numerous benefits of their use, are topics widely treated in the scientific literature. However, the aspects related to the calculation of the scores that derive from its use, as well as the reliability and validity of the rubrics have received less attention. The present work, therefore, focuses on the exposition of a set of methodological limitations related to the rubrics and the proposal of feasible solutions to overcome them. The final considerations of the work focus on proposals related to rubrics where the scores are not equivalent to the grading system of the reference country, as well as to its weighting. We also emphasize the biases that derive from the use of correlation and reliability coefficients in the use of intra- and inter-observer reliability and in the validity of content and the desirability to replace with agreement coefficients.

Keywords: Rubrics. Skills assessment. Reliability & Validity.

Limitações metodológicas e soluções viáveis na avaliação e cálculo das qualificações obtidas através das rubricas como estratégias de avaliação de competências de estudantes universitários e outros níveis

Resumo

O desenho e elaboração de rubricas como estratégias de avaliação no ensino superior e outros níveis educacionais, bem como os inúmeros benefícios de seu uso, são temas amplamente tratados na literatura científica. No entanto, os aspectos relacionados ao cálculo dos escores que derivam do seu uso, bem como a confiabilidade e validade das rubricas receberam menos atenção. O presente trabalho, portanto, concentra-se na exposição de um conjunto de limitações metodológicas relacionadas às rubricas e à

proposta de soluções viáveis para superá-las. As considerações finais do trabalho se concentram em propostas relacionadas a rubricas onde as pontuações não são equivalentes ao sistema de classificação do país de referência, bem como à sua ponderação. Os viés que derivam do uso de coeficientes de correlação e confiabilidade no uso da confiabilidade intra e inter-observador e na validade do conteúdo e a conveniência de sua substituição por coeficientes de concordância também são destacados.

Palavras-chave: Rubricas. Avaliação de competências. Fiabilidade e validade.