Muro de la Investigación, 2022(2), julio-diciembre
ISSN: 2523-2886

Propuesta de una rúbrica para evaluar la calidad de las tesis doctorales: Un enfoque de evaluación formativa

Proposal of a rubric to evaluate the quality of doctoral theses: A formative evaluation approach

Jaime Natanael Gonzales Lopez[1a]

Universidad Peruana Unión, Lima, Perú[1]

Recibido: 15 de agosto de 2020 Aceptado**:** 05 de enero de 2021

Resumen

La presente investigación tuvo como objetivo determinar las propiedades psicométricas de una rúbrica diseñada, con fines formativos, para evaluar la calidad de las tesis doctorales en programas de PhD en Educación. La rúbrica está constituida por seis dimensiones, y sus puntuaciones van de 1 a 4. Para determinar su grado de validez se usó el coeficiente V de Aiken, y para determinar su nivel de fiabilidad se usaron el coeficiente de concordancia Kappa de Cohen y el coeficiente de concordancia Kappa de Fleiss. En ambos procedimientos participaron 6 investigadores con grado de PhD en Educación como jueces y evaluadores. Además, se seleccionó una tesis de cada una de las 31 mejores universidades de mundo en el área de educación para realizar el análisis de fiabilidad inter-observador. Luego de realizar los análisis respectivos, se concluyó que cada dimensión de la rúbrica tiene un alto nivel de validez al obtener puntuaciones superiores a 0.83; y que la rúbrica es altamente fiable puesto que la fuerza de concordancia entre uno y otro observador, así como en su conjunto, resultaron ser considerables o casi perfectas.

Palabras clave: Investigación, tesis doctorales, rigurosidad científica, calidad, rubrica, educación

Abstract

The objective of this research was to determine the psychometric properties of a rubric designed, with formatives purposes, to evaluate the quality of doctoral theses in PhD programs in Education. The rubric is made up of six dimensions, and its scores range is from 1 to 4. To determine its degree of validity, the Aiken V coefficient was used, and to determine its level of reliability, Cohen's Kappa concordance coefficient and the coefficient of Kappa

[a]Correspondencia al autor:
E-mail: natanael@upeu.edu.pe
Este trabajo esta basado en un trabajo de tesis: http://hdl.handle.net/20.500.12840/4367

de Fleiss were used. In both procedures, 6 researchers with a PhD in Education participated as judges and evaluators. In addition, a thesis was selected from each of the 31 best universities in the world in the area of education to perform inter-observer reliability analysis. After of the respective analyzes, it was concluded that each dimension of the rubric has a high level of validity when obtaining scores higher than 0.83; and that the rubric is highly reliable since the strength of agreement between one and another observer, as well as a whole, turned out to be considerable or almost perfect.

Keywords: Research, doctoral theses, scientific rigour, quality, rubric, education

Introducción

Las universidades y escuelas de posgrado ayudan a crear economías más competitivas a nivel local, regional y de todo el globo, en especial, los programas doctorales. Por ese motivo, en los países aún no desarrollados, existe la imperativa la necesidad de mejorar la calidad de enseñanza e investigación de los programas doctorales, a través de la construcción y creación de conocimientos nuevos y avanzados; y de rediseñar los programas doctorales para crear instituciones superiores de rango mundial que puedan competir en el vasto mercado educativo mundial (Salmi, 2009).

En este sentido, ser una universidad de rango mundial significa ser una universidad de investigación. La universidad de investigación se encarga de formar a un grupo minoritario de estudiantes doctorales, que son generalmente los mejores y los más brillantes del país, contratando a los académicos mejor cualificados. Ellos, aunque son los encargados de producir la mayor parte de los descubrimientos científicos, son claves en la formación de los futuros doctores que generarán aportes sustanciales al conocimiento (Altbach & Salmi, 2011).Tales aportes empiezan a realizarse en la tesis doctoral, y continúan luego de la obtención del grado. No obstante, siendo la tesis doctoral un elemento tan importante para el progreso del conocimiento, aún muchas universidades no comprenden a cabalidad lo que implica llevar a cabo tal desafío, especialmente las universidades que no son de rango mundial.

Por ello, es importante destacar que la tesis doctoral es el reflejo de las capacidades alcanzadas por el candidato a doctor, y permite conocer cuán preparado se encuentra para realizar investigaciones rigurosas, novedosas, originales y de alto impacto. Sin embargo, las diferentes maneras de pensar de los asesores y dictaminadores sobre lo que debe y no debe

hacerse en una tesis doctoral, ha llevado a que muchos candidatos se desanimen en el camino o prefieran hacer algo sin relevancia académica.

Las divergencias de pensamiento, incluso, llegan a influir en la calificación que se le da a la tesis; y lamentablemente, en países en los cuáles no hay universidades de investigación, la realidad es más crítica, ya que existen muchas limitantes a nivel teórico y metodológico. Es decir, tesis que en universidades de investigación podrían ser señaladas como originales y significativas, son rechazas solo porque el asesor y los dictaminadores tienen limitaciones al momento de comprender tales contribuciones; y, por el contrario, existe cierta cantidad de tesis doctorales que no presentan nada original ni significativo, pero son aprobadas con altas calificaciones. Esta inconsistencia de criterios, de igual forma, está llevando a generar una brecha importante entre lo producido por las universidades de investigación, y lo producido por las demás universidades.

En la mayor parte de los casos, son las universidades, facultades o gobiernos los encargados en definir los criterios para juzgar la calidad de las tesis doctorales; no obstante, aún no existe consenso, global o por disciplina académica, sobre los criterios a tomarse en cuenta (Walker et al., 2008). Esto da como resultado una situación paradójica, pues al ser la tesis doctoral un elemento objetivo se la evalúa con criterios puramente subjetivos.

Según algunos estudios que analizan los criterios de evaluación de cada uno de los miembros del comité dictaminador, se descubrió que, inclusive cuando una universidad establece criterios para evaluar la calidad de una tesis doctoral, los dictaminadores experimentados se guían más por sus propios juicios, y no por los criterios de la institución universitaria, al momento de juzgar si una tesis doctoral cumple o no con los estándares de calidad requeridos. En cambio, los dictaminadores inexpertos se guían más por los criterios establecidos por la universidad, pero no llegan a establecer con claridad los límites al momento de juzgar como buena o mala una tesis doctoral, y ven su incertidumbre como un problema serio, en especial cuando no se encuentran familiarizados con el tema de tesis (Mullins & Kiley, 2002; Kiley & Mullins, 2004).

Instituciones como el Council of Graduate Schools (2005) y la European University Association (2005) han establecido lineamientos generales sobre la importancia, propósito y calidad de una tesis doctoral. También se han llevado a cabo algunas investigaciones relevantes (Isaac, Quinlan, & Walker, 1992; Adams & White, 1994; Lovitts, 2006), desde

finales del siglo XX, que han tratado de establecer criterios para la evaluación de la calidad de las tesis doctorales. Dichos estudios se han centrado principalmente en establecer criterios generales para todas las ciencias, o para ciertas ciencias específicas como sociología, economía, administración, psicología, entre otras. Sin embargo, no se pudo encontrar algún estudio que se centre en establecer criterios para evaluar la calidad de las tesis doctorales en educación, generando una brecha de conocimiento que precisa ser cubierta. Por tal motivo, se determinará las propiedades psicométricas de la rúbrica propuesta en esta investigación para evaluar la calidad de las tesis doctorales en educación.

Materiales y métodos

La construcción y el análisis de las propiedades psicométricas de una rúbrica pertenecen al campo de la psicometría. Los investigadores construyen una rúbrica cuando desean evaluar, por medio de descriptores, el nivel de calidad de desempeño de una actividad. Construir una rúbrica no solo implica la reunión de criterios para medir algún constructo, sino demanda que el investigador lleve a cabo un metódico y minucioso estudio que permita tener una rúbrica confiable y válida (Stevens & Levi, 2005; DeVellis, 2012; Furr, 2011; Spector, 1992). En el presente estudio se usó el diseño transversal, dado que las tesis evaluadas para analizar la fiabilidad de la rúbrica propuesta son seleccionadas en un solo momento en el tiempo (Little, 2013; Mills & Gay, 2016).

Participantes

Para realizar el análisis de fiabilidad se escogió como población a todas las tesis doctorales producidas por las mejores universidades de rango mundial para estudiar educación, a nivel de pregrado y/o posgrado. Estas instituciones fueron seleccionadas tomando en cuenta los dos rankings académicos de universidades más reconocidas internacionalmente del 2018. El primero de ellos es el ranking publicado por la Universidad Jiao Tong de Shanghái (ARWU); y el segundo es el ranking publicado por el Times Higher Education (THE). Ambos rankings cuentan no solo con una clasificación mundial global, sino también cuentan con una clasificación mundial por cada disciplina académica.

Para esta investigación, solo se tomó en cuenta la clasificación que permite conocer las mejores universidades en el área de educación. Ambos rankings presentaron ciertas

diferencias al momento de posicionar las mejores instituciones a nivel mundial en el área de educación, por ello se consideró incluir en la investigación solo a aquellas con cumplían con los siguientes requisitos: que se encuentren presentes dentro de las 50 mejores universidades en ambos rankings; que cuenten con un Programa Doctoral Académico (PhD) en Educación; y que las tesis estén disponibles en inglés.

Por ello, al considerar el primer requisito, se observó que eran 31 las universidades que se encontraban en ambos rankings: 23 de Estados Unidos, 2 de Australia, 2 de Reino Unido, 1 de Canadá, 1 de Países Bajos, 1 de Singapur y 1 de China. Al considerar el segundo requisito, se vio que las 31 universidades contaban con al menos un programa de PhD en Educación. Es resaltable el hecho de que en el ranking de las mejores escuelas de posgrado de Estados Unidos, publicado por la US News, hayan aparecido 22 de las 23 instituciones americanas señaladas con anterioridad dentro de las 50 mejores instituciones americanas donde uno puede seguir un programa de PhD en Educación, a excepción de Florida State University que aparece en el puesto 52; no obstante, se la consideró al estar presente entre las 50 mejores instituciones en el área de educación, según el ARWU y el THE. Al considerar el tercer requisito, también se pudo visualizar que cada una de las 31 universidades contaban con tesis doctorales en inglés. Por lo tanto, la población quedó conformada por las 31 universidades que cumplieron con los tres criterios de inclusión y exclusión.

Al ser el tamaño de la población de tesis infinita y desconocida se consideró usar un muestreo intencional para seleccionar las tesis doctorales que participarían en el estudio; por ello, se eligió una tesis doctoral aleatoriamente de cada una de las 31 universidades participantes. A fin de llevar a cabo la validez de contenido y la fiabilidad inter-observador se solicitó a 6 investigadores con grado de PhD en educación que actuaran en calidad de jueces y evaluadores. Todos estos especialistas son docentes en escuelas de posgrado, y cuenta con una gran experiencia como asesores y dictaminadores a nivel doctoral en Estados Unidos (3), Reino Unido (2) y Australia (1).

Instrumentos

La rúbrica diseñada para evaluar la calidad de las tesis doctorales en educación se encuentra divida en seis dimensiones: introducción, revisión literaria, teoría, metodología, resultados,

y discusión y conclusión. La puntuación de cada dimensión va desde inaceptable (1) hasta sobresaliente (4). Esta rúbrica fue diseñada para interpretar que, a mayor puntaje, mayor es la calidad de la tesis doctoral, y fue elaborada tomando en consideración los estudios hechos por Lovitts (2006) y por Boote y Beile (2005).

Aunque la rúbrica puede servir para evaluar una tesis doctoral ya sustentada o para la dictaminación final de la misma; en realidad, fue elaborada con un fin formativo; es decir, se desea que esta rúbrica sirva como una guía para que el estudiante conozca lo que se espera de su tesis doctoral. Por esta razón, la presente rúbrica puede ser utilizado durante todo el tiempo que el estudiante tome el programa doctoral o elabore su tesis.

Se eligió construir una rúbrica y no una lista de cotejo o una escala debido a que estas dos últimas carecen de descripciones de la calidad del desempeño. El uso de una lista de cotejo es excelente cuando se necesita saber si se ha hecho algo o no, y las escalas son excelentes cuando se quiere conocer, por ejemplo, la frecuencia de un constructo; pero, las rúbricas permiten describir cada nivel que se está evaluando, logrando que el proceso de evaluación sea más comprensible (Brookhart, 2013).

Análisis de datos

En la primera etapa, para evaluar la validez de contenido de la rúbrica diseñada, los seis doctores actuaron como jueces; así se pudo definir mejor los criterios e ítems de cada dimensión. Posteriormente, con la rúbrica ya válida, ellos mismos, actuaron como evaluadores de las 31 tesis doctorales (PhD) seleccionadas como muestra, para conocer si la rúbrica es fiable o no. Para analizar las propiedades psicométricas de la rúbrica se usaron los softwares estadísticos SPSS 24 y R Project. Para cuantificar la validez de contenido por criterio de jueces se usó el coeficiente V de Aiken; para ello, cada juez evaluó cada dimensión en cuatro categorías, con puntuaciones de 0 a 3 (ver tabla 1).

Tabla 1
Categorías para la calificación del coeficiente V de Aiken

Categoría	Calificación	Indicador
Claridad Cada ítem de la dimensión se comprende fácilmente, es decir, su semántica y sintaxis son adecuadas	0 = No cumple el criterio	Los ítems no son claros
	1 = Bajo nivel	La mayoría de ítems necesitan modificaciones semánticas y sintácticas
	2 = Moderado nivel	Algunos ítems necesitan modificaciones semánticas y sintácticas
	3 = Alto nivel	Los ítems son claros, y están bien redactados a nivel semántico y sintáctico
Coherencia Cada ítem tiene relación lógica con la dimensión	0 = No cumple el criterio	Los ítems no tienen ninguna relación lógica con la dimensión
	1 = Bajo nivel	Los ítems tienen una relación tangencial con la dimensión
	2 = Moderado nivel	Los ítems tienen una relación moderada con la dimensión
	3 = Alto nivel	Los ítems se encuentran completamente relacionados con la dimensión
Relevancia Cada ítem es esencial o importante	0 = No cumple el criterio	Casi todos los ítems pueden ser eliminados sin que se afecte la medición de la dimensión
	1 = Bajo nivel	Existe duplicidad de ítems, o un ítem puede estar incluido en otro ítem
	2 = Moderado nivel	La mayoría de ítems son relevantes
	3 = Alto nivel	Todos los ítems son relevantes y deben permanecer en la dimensión
Suficiencia Cada uno de los ítems que pertenece a una misma dimensión bastan para obtener la medición de ésta.	0 = No cumple el criterio	Los ítems son insuficientes para medir la dimensión
	1 = Bajo nivel	Los ítems miden algún aspecto de la dimensión, pero no toda la dimensión
	2 = Moderado nivel	Se deben añadir algunos ítems para así poder evaluar la dimensión por completo
	3 = Alto nivel	Los ítems son suficientes para medir la dimensión

Para determinar si la rúbrica es confiable se realizó el análisis de fiabilidad inter-observador por medio del coeficiente de concordancia Kappa de Cohen y el coeficiente de concordancia Kappa de Fleiss. Los observadores evaluaron, en primera instancia, cada dimensión de la tesis dándole una puntuación de 1 a 4, donde 1 significa inaceptable, 2 aceptable, 3 muy buena y 4 sobresaliente. Luego, sumaron los puntajes obtenidos de las seis dimensiones y le dieron una puntuación global a cada tesis, con puntuaciones de 1 a 4

también (ver tabla 2). Los puntos de corte de las puntuaciones de la rúbrica fueron estimados siguiendo las sugerencias de los observadores.

Tabla 2
Puntos de corte de rúbrica propuesta

Nivel de uso	Puntuación	Valor
Sobresaliente	22 - 24	4
Muy Buena	16 - 21	3
Aceptable	10 - 15	2
Inaceptable	6 - 9	1

Resultados y discusión

En primer lugar, se presentará los resultados del análisis de validez de contenido por criterio de jueces, y luego se procederá a mostrar los resultados del análisis de fiabilidad por concordancia de la rúbrica.

Análisis de validez de contenido por criterio de jueces

Coeficiente de validez V de Aiken

Para encontrar el coeficiente de validez V de Aiken de cada dimensión fue necesario sumar cada uno de los puntajes otorgados por los evaluadores, y dividirlo entre el número de evaluadores multiplicado por el número de valores asignados menos 1. En el caso del presente estudio se consideró cuatro valores: 3 para calificar como alto nivel de validez, 2 para calificar como moderado nivel de validez, 1 para calificar como bajo nivel de validez, 0 para calificar como nulo nivel de validez.

Al realizar el análisis de validez de contenido, los jueces evaluaron las categorías de claridad, coherencia, relevancia y suficiencia de los ítems que integraban cada una de las dimensiones. Para que la dimensión fuese juzgada como válida los valores de cada categoría debieron ser iguales o mayores a 0.83, así que si alguna categoría obtenía un valor menor se procedería a levantar las observaciones según la sugerencia dada por los jueces. Las dimensiones teoría, y discusión y conclusiones no presentaron ninguna observación, por la cual, no se realizaron modificaciones; en cambio, las dimensiones introducción, revisión literaria, metodología y resultados presentaron observaciones en alguna categoría, por ello,

se procedió a cambiar o corregir algunos ítems de las dimensiones según las sugerencias hechas por los jueces.

Análisis de fiabilidad por concordancia

Coeficiente de concordancia Kappa de Cohen

Después de validar la rúbrica, se procedió a analizar la fiabilidad de la misma por medio del coeficiente de concordancia Kappa de Cohen, tomando en cuenta las siguientes consideraciones: Si el "valor p" resultaba ser menor a 0.05, se rechazaba la hipótesis nula (Ho), y se aceptaba la hipótesis alternativa (H1); pero si era mayor a 0.05, se aceptaba la hipótesis nula. Cabe señalar que la hipótesis nula señalaba que no existía concordancia entre los resultados de los observadores, y la hipótesis alternativa que sí existía dicha concordancia. Luego de realizar el análisis de Kappa de Cohen de dos en dos se obtuvo una fuerza de concordancia entre considerable y casi perfecta.

Coeficiente de concordancia Kappa de Fleiss

Luego de haber realizado el análisis de fiabilidad por medio del coeficiente de concordancia Kappa de Cohen, se pudo observar que la rúbrica es altamente fiable. No obstante, el Kappa de Cohen tiene una limitación, sólo puede medir el nivel de concordancia entre dos observadores; por ello, se procedió usar también el coeficiente Kappa de Fleiss para analizar el nivel de concordancia en conjunto de todos los observadores.

En la tabla 3, se aprecia que el coeficiente de concordancia Kappa de Fleiss fue de .804 y el valor p = .000. A un nivel de significancia de .05, se cumple que $p < \alpha$, lo que significa que se rechaza la hipótesis nula y se acepta la hipótesis alternativa. Por lo tanto, existe concordancia entre los resultados de los seis observadores. Además, el resultado del coeficiente de concordancia indica que existe una fuerza de concordancia casi perfecta entre los seis observadores, lo que demuestra una vez más la eficacia de la rúbrica en la evaluación de calidad de las tesis doctorales en educación.

Tabla 3
Análisis de fiabilidad por concordancia entre los seis observadores

	Valor	Error estándar asintótico	Z	Significación aproximada
Kappa de Fleiss	.804	.044	18.192	.000

Conclusión

La evaluación es un elemento esencial al momento de poder medir el nivel de los estudiantes doctorales, así como la calidad de sus trabajos de investigación para obtener el grado de doctor (Maki & Borkowski, 2006). Como se ha visto, en las últimas dos décadas ha cobrado impulso la evaluación de la educación doctoral, así como la evaluación de las tesis doctorales, y parece que esto irá ganado fuerza con el tiempo; por ello, se ha realizado, y continúa llevándose a cabo, una serie de iniciativas que permitan mejorar la calidad de las tesis doctorales, como este estudio, que pretende contribuir a esta mejora con una rúbrica válida y fiable.

De acuerdo con los resultados que fueron presentados en esta investigación sobre las propiedades psicométricas de la rúbrica diseñada para evaluar la calidad de las tesis doctorales en educación, se concluye que la rúbrica propuesta es válida y fiable para evaluar la calidad de las tesis doctorales en educación. En este sentido, es recomendable que los directores de escuelas de posgrados en educación deberían convocar a dictaminadores y asesores para discutir y crear estándares para evaluar la calidad de las tesis doctorales. La universidad podría luego componer sus propios estándares de calidad, de la misma manera que tienen declaraciones sobre la estructura y naturaleza de la tesis. La existencia de tales estándares también permitiría a las universidades juzgar las afirmaciones de los estudiantes de que los docentes de investigación no son aptos o están desactualizados, o que su tesis no fue juzgada de manera justa.

Referencias

Adams, G. B., & White, J. D. (1994). Dissertation Research in Public Administration and Cognate Fields: An Assessment of Methods and Quality. *Public Administration Review*, *54* (6), 565–576. http://www.jstor.org/stable/976677

Altbach, P. G., & Salmi, J. (Eds.). (2011). *The Road to Academic Excellence-The Making of World-Class Research universities*. Washington, DC: The World Bank.

Boote, D., & Beile, P. (2005). Scholars Before Researchers: On the Centrality of the Dissertation Literature Review in Research Preparation. *Educational Researcher*, *34*(6), 3–15. http://edr.sagepub.com/cgi/doi/10.3102/001318 9X034006003

Brookhart, S. (2013). *How to Create and Use Rubrics for Formative Assessment and Grading*. Alexandria: ASCD.

Council of Graduate Schools. (2005). *The Doctor of Philosophy Degree: A Policy Statement*. Washington, DC: Council of Graduate Schools.

DeVellis, R. (2012). *Scale Development: Theory and Applications* (3rd ed.). Washington DC: SAGE Publications.

European University Association. (2005). *Doctoral Programmes for the European knowledge Society. Report on the EUA Doctoral Programmes Project, 2004–2005*. Brussels: European University Association.

Furr, R. (2011). *Scale Construction and Psychometrics for Social and Personality Psychology*. Thousand Oaks: SAGE Publications.

Isaac, P., Quinlan, S., & Walker, M. (1992). Faculty Perceptions of the Doctoral Dissertation. *The Journal of Higher Education*, *63* (3), 241–268. http://www.jstor.org/stable/1982014

Kiley, M., & Mullins, G. (2004). Examining the examiners: How inexperienced examiners approach the assessment of research theses. *International Journal of Educational Research*, *41* (2), 121–135. http://ezproxy.concytec.gob.pe:2053/science/article/pii/S0883035505000224

Little, T. D. (Ed.). (2013). *The Oxford Handbook of Quantitative Methods: Foundations*. New York: Oxford University Press.

Lovitts, B. (2006). Making the Implicit Explicit. In P. Maki & N. Borkowski (Eds.), *The Assessment of Doctoral Education: Emerging Criteria and New Models for Improving Outcomes*. Sterling, Virginia: Stylus Publishing.

Mills, G., & Gay, L. (2016). *Educational Research: Competencies for Analysis and Applications* (11th ed.). Boston: Pearson.

Mullins, G., & Kiley, M. (2002). It's a PhD, not a Nobel Prize: How experienced examiners assess research theses. *Studies in Higher Education, 27* (4), 369–386. https://www.uow.edu.au/content/groups/public/@web/@raid/documents/doc/uow016364.pdf

Salmi, J. (2009). *The Challenge of Establishing World-Class Universities*. Washington, DC: The World Bank.

Spector, P. (1992). *Summated Rating Scale Construction: An Introduction*. Newbury Park: SAGE Publications.

Stevens, D., & Levi, A. (2005). *Introduction to Rubrics: An Assessment Tool to Save Grading Time, Convey Effective Feedback and Promote Student Learning*. Sterling, Virginia: Stylus Publishing.

Walker, G., Golde, C., Jones, L., Conklin, A., & Hutchings, P. (2008). *The Formation of Scholars: Rethinking Doctoral Education for the Twenty-First Century*. San Francisco: Jossey-Bass.