XVI Congreso Nacional de Investigación Educativa
CNIE-2021

Evidencias de validez de las consecuencias de la Rúbrica de Evaluación del Desempeño Docente

María del Rayo López Contreras
a310028@uabc.edu.mx

Área temática 12. Evaluación educativa.
Línea temática: Desarrollo y validación de instrumentos.
Porcentaje de avance: 60%.
Programa de posgrado: Doctorado en Ciencias Educativas Sexto cuatrimestre.
Institución donde realiza los estudios de posgrado: Instituto de Investigación y Desarrollo Educativo.

Resumen

La meta del presente trabajo es aportar evidencias de validez de las consecuencias de la Rúbrica de Evaluación del Desempeño Docente (REDD, la cual tiene el objetivo de evaluar la práctica pedagógica para ofrecer retroalimentación al desempeño del docente, y su uso es formativo de docentes. El método está dividido en dos estudios: cuantitativo y cuasi-experimental, a través de un diseño de línea base múltiple por participantes; cualitativo, interpretativo y método de estudio de caso. Para ambos estudios las participantes son alumnas de séptimo y octavo semestre de la Licenciatura en Educación Preescolar que están realizando sus prácticas; así como una docente experta en la evaluación de docentes en formación. Tras establecer la línea base (mínimo tres sesiones), se retroalimenta a las practicantes sobre su desempeño; posteriormente se califican y retroalimentan tres sesiones más; al finalizar se entrevistarán a las participantes. Se pretende identificar cambios en el desempeño de la práctica, así como conocer la experiencia y significado de las estudiantes sobre el proceso de retroalimentación y de la experta sobre el uso de la rúbrica, a la vez que se identifican las consecuencias de la misma. Con la investigación se busca aportar al estado del conocimiento del diseño y validación de instrumentos en cuanto a la validez de los usos y consecuencias en instrumentos con uso formativo.

Palabras clave: *Validez de las consecuencias, evaluación formativa, práctica docente, educación preescolar.*

Introducción

Las aseveraciones sobre los resultados de los instrumentos de evaluación educativa deberían estar sustentados en evidencias de *validez*, cuya importancia es fundamental al desarrollar un instrumento. La validez es un juicio evaluativo global del grado en el que la evidencia empírica y los fundamentos teóricos apoyan la idoneidad y adecuación de las inferencias y acciones, basadas en los puntajes de las pruebas u otros modos de evaluación. Dichas inferencias deben ser enfocadas para los objetivos y usos propuestos de las pruebas (Messick, 1988, como se citó en Messick, 1989, American Educational Research Association [AERA], American Psychological Association [APA] y la National Council on Measurement in Education [NCME], 2014).

La AERA, APA y la NCME (2014) hacen referencia a las *evidencias de validez*, que pueden señalar diferentes aspectos de la misma pero no representan distintos tipos de validez. Entre esas evidencias se encuentran las que se refieren a las consecuencias, y se dividen en: las que proceden directamente de las interpretaciones y usos de las puntuaciones de los instrumentos previstas por los desarrolladores; las que son reclamaciones sobre los usos que no están directamente basados en las interpretaciones de las puntuaciones; y las que son consecuencias involuntarias.

Acorde a lo anterior, Pedroza y Luna (2017) desarrolaron la Rúbrica de Evaluación del Desempeño Docente (REDD) la cual tiene el objetivo de evaluar la práctica pedagógica para ofrecer retroalimentación al desempeño del docente, y su uso es formativo de docentes. La rúbrica está alineada al Programa de Educación Preescolar 2017, dividida en tres grandes dimensiones: (a) planeación, (b) intervención y (c) evaluación; en una escala de intervalo de cuatro puntos (Insatisfactorio, En proceso, Satisfactorio y Experto). Se evalúa el desempeño docente a través de la observación de la interacción en el aula durante una videograbación de una sesión de 18 a 20 minutos de duración, el análisis de la planeación y de las evidencias de evaluación del aprendizaje de los niños. Con base en los resultados de la REDD se genera un reporte individualizado con el nivel de cada reactivo así como su descripción, se entrega a la educadora durante la entrevista de retroalimentación realizada de forma individual, privilegiando el diálogo entre el evaluador y el evaluado. En su primera versión, cuenta con evidencias de validez de consistencia interna, de contenido y del proceso de respuesta (Pedroza y Luna, 2017)

El presente tiene como objetivo general aportar evidencias de validez de las consecuencias para el uso formativo del instrumento. A continuación se presentan los objetivos específicos:

Objetivos Específicos

1. Identificar los cambios en el desempeño de las practicantes evaluadas tras la retroalimentación dada con base en los resultados de la REDD
2. Reconocer los significados y experiencias del evaluador y evaluado sobre la retroalimentación ofrecida con la REDD
3. Identificar consecuencias no previstas del uso de la REDD.

Preguntas de investigación

1. ¿La retroalimentación dada con base en los resultados de la REDD genera un cambio en el desempeño de las participantes?
2. ¿Los significados y experiencias del evaluador y evaluado reflejarán un sentido positivo y de mejoramiento tras la retroalimentación?
3. ¿Los significados y experiencias de la experta reflejarán un sentido positivo del uso de la REDD?
4. ¿Existen consecuencias no previstas de la REDD?

Hipótesis

1. La retroalimentación dada con base en los resultados de la REDD generará cambios en el desempeño de las practicantes.

No obstante la importancia de aportar evidencias de validez de los instrumentos, particularmente la de las consecuencias, son pocos las investigaciones que la reportan. Collie y Zumbo (2014) y Shear y Zumbo (2014) realizaron estudios donde examinaron las prácticas de validación o estudios de validez presentadas en varios artículos publicados, en dos décadas diferentes. En la revista *Journal of Educational Psychology*, todo lo publicado en la década del 2000 al 2010; y de la revista *Educational and Psychological Measurement*, en las décadas de 1960-69 y del 2000-2009. En ambos estudios se encontraron diferentes aportes de evidencias de validez, mayoritariamente las referidas a la estructura interna y de relación con otras variables, pocas a las del proceso de respuesta y nulas a las referidas a la validez de las consecuencias.

La autora de esta investigación realizó el inicio de una revisión sistemática de la literatura sobre validez de las consecuencias en evaluación formativa, cuyo objetivo es identificar las aportaciones de evidencias de validez de las consecuencias de instrumentos de evaluación formativa. La combinación de palabras clave fue: *validity/ consequential/ formative/ assessment*; con los criterios de inclusión: artículos arbitrados de la base de datos ERIC, en inglés o español, de últimos 10 años (2012 a 2021), del área de conocimiento educación, investigación educativa, evaluación educativa, dentro de *test validity.*

Se encontró la inexistencia de artículos en español, en inglés un total de 75, de los cuales 37 mencionaban validez de las consecuencias de evaluación en el resumen; 22 son empíricos, diez teóricos y cinco aun no identificados. Cabe resaltar que aún falta el análisis de cada artículo, posterior a este se deberán eliminar los que no aporten evidencia de validez de las consecuencias de uno o varios instrumentos.

Con los estudios realizados por Collie y Zumbo (2014) y Shear y Zumbo (2014), aunado al primer acercamiento de la autora, queda patente el reducido número de artículos sobre evidencias de validez de las consecuencias en la evaluación formativa, lo cual podría reflejarse también en el estado del conocimiento del desarrollo y validación de instrumentos.

Desarrollo

Marco teórico

Existen diferentes organismos que instauran normas para regular la adecuada elaboración y práctica de los instrumentos educativos y psicológicos, entre los cuales se encuentran la AERA, APA y NCME. Sus estándares tienen como objetivo brindar criterios para el desarrollo y evaluación de pruebas o instrumentos y sus prácticas, además, proporcionar pautas para evaluar la validez de las interpretaciones de los puntajes obtenidos en las pruebas y los usos previstos (AERA, APA y NCME, 2014; Kane y Staiger, 2013).

Dentro de los estándares mencionados con anterioridad se señala que todo instrumento de evaluación educativa y psicológica debe brindar evidencias de *validez*, concepto unificado o unitario basado en diferentes evidencias que se aportan para corroborar el constructo que se busca evaluar, así como las interpretaciones basadas en los resultados de dicha evaluación. A lo anterior, se agrega que las inferencias deben ser enfocadas para los objetivos y usos propuestos de las pruebas; y que es la cuestión más importante y fundamental al desarrollar un instrumento: Finalmente existen diferentes formas de aportar evidencias de validez (Messick, 1988, como se citó en Messick, 1989, AERA, APA y, NCME, 2014).

Acorde con el concepto de validez, la AERA, APA y la NCME (2014) hacen referencia a las *evidencias de validez*, que pueden señalar diferentes aspectos de la misma pero no representan distintos tipos de validez. Dichas evidencias son las basadas en el contenido, en el proceso de respuesta, en la estructura interna, en las relaciones con otras variables, y las evidencias de las consecuencias. Éstas últimas se dividen en: las que proceden directamente de las interpretaciones y usos de las puntuaciones de los instrumentos previstas por los desarrolladores; las que son reclamaciones sobre los usos que no están directamente basados en las interpretaciones de las puntuaciones; y las que son consecuencias involuntarias.

Es necesario recordar que la rúbrica tiene el objetivo de evaluar la práctica pedagógica para ofrecer retroalimentación al desempeño del docente, y su uso la evaluación formativa de docentes.

Así pues, enlazando parte de la teoría que fundamenta a la REDD (la evaluación formativa, EF) con las evidencias de validez de las consecuencias, La EF, en general, trata sobre brindar información, dar retroalimentación y buscar generar un cambio. Tejedor (2012) explica que el sentido formativo de la evaluación reside en el supuesto de pensar que la información (retroalimentación) dada al profesor va a estimularlo a efectuar cambios oportunos. Es necesario recordar que la REDD tiene como uso la evaluación formativa de docentes, es decir, con base en los resultados de la rúbrica se brinda retroalimentación a la educadora con la finalidad de mejorar su práctica, consecuencia del uso de la rúbrica. Con base en lo anterior, se presenta la necesidad de identificar el grado en que la retroalimentación dada con base en los resultados de la REDD aporta a la mejora del desempeño docente; cumpliendo así con los estándares de calidad de la validación y desarrollo de un instrumento de la AERA, APA y NCME.

Método

Está dividido en dos estudios: La calificación del desempeño docente y la retroalimentación a las practicantes por parte de una docente de práctica profesional y; la identificación de los significados y experiencias de las practicantes sobre el proceso de retroalimentación y de la experta sobre el uso de la rúbrica.

Cabe resaltar que para ambos estudios las participantes serán cinco alumnas de séptimo y octavo semestre, que se encuentren realizando sus prácticas profesionales frente a grupo para CONAFE. Las estudiantes están realizando la Licenciatura en Educación Preescolar, de la Benemérita Escuela Normal Estatal, profesor Jesús Prado. Así como una docente de práctica profesional, con dos años de experiencia frente aula y 17 como docente de la misma Normal Estatal, así como en el uso de la REDD.

Estudio 1. Se basará en un enfoque cuantitativo y cuasi-experimental con un diseño de línea base múltiple a través de participantes. En los diseños de línea base múltiple se demuestra el efecto de una contingencia mostrando que un cambio de comportamiento se relaciona con la introducción de la misma en diferentes puntos en el tiempo, es decir, muestra la eficacia de un tratamiento/intervención sobre un cambio en la conducta. El efecto del tratamiento/intervención se comprueba cuando se presenta un patrón de cambio conforme se introduce la intervención. En particular, los diseños de base múltiple a través de participantes se aplica el mismo tratamiento en series, a la misma conducta de distintos individuos en el mismo ambiente (Kazdin, 2000; Kerlinger, 2008).

Los instrumentos y materiales a utilizar son:

- REDD. Alineada al Programa de Educación Preescolar 2017, dividida en tres grandes dimensiones: (a) planeación, (b) intervención y (c) evaluación; en una escala de intervalo de cuatro puntos (Insatisfactorio, En proceso, Satisfactorio y Experto). En la primer versión (Pedroza, 2016) se obtuvo la confiabilidad y se aportó diferentes evidencias de validez. Para la primera, el alfa de Cronbach y la ordinal, así como Theta de Armor y Coeficiente Theta ordinal, oscilaron entre.85 a .91. En cuanto a las evidencias de validez se obtuvieron las de contenido, de la estructura interna y las del proceso de respuesta. De contenido se obtuvo a través de jueceo de un grupo de expertos, análisis de los documentos normativos y la documentación del proceso de construcción del instrumento. Con el AFE se obtuvo el índice de bondad de ajuste GFI de.739 y RMSR de .063; y Alfa de Cronbach de .85 a .89 en los tres factores. El porcentaje de acuerdo exacto entre los observadores fue Alto (53% a 91%). Los coeficientes Kappa de Cohen en su versión cuadrático ponderado cayeron en Aceptable, moderada y considerable confiabilidad. El coeficiente de Generalizabilidad se identificaron los coeficientes para planeación: .63, intervención: .75, y evaluación: .85.
- Protocolo de retroalimentación. Dividido en siete orientaciones: establecer un plan de mejora, seleccionar aspectos específicos para retroalimentar, ser consistente con los criterios de evaluación establecidos, establecer un diálogo bidireccional con quien se retroalimenta, establecer un clima de confianza, promover la autorregulación y entregar informe por escrito y confidencial. Se realizó con base en la Revisión Sistemática de la Literatura de Pedroza y García Poyato (en prensa).

- Tablets, memorias externas de 64 GB.

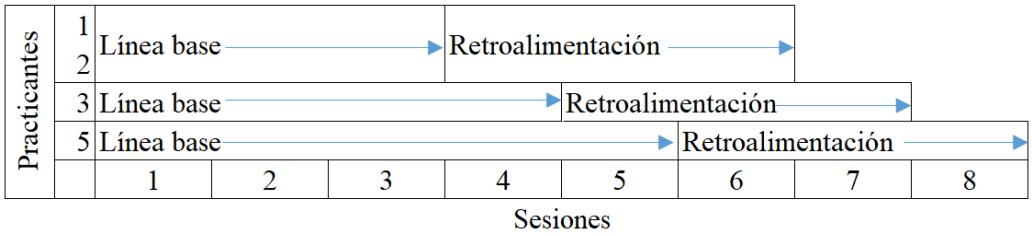
Procedimiento

Se aplica el diseño de línea base múltiple *a través de participantes*, en este diseño se necesita identificar la conducta a medir (práctica docente), la intervención o tratamiento (retroalimentación) y el ambiente en el que se medirá (aula de preescolar). La entrega de las evidencias de las practicantes es mediante un portafolio integrado por la videograbación de la sesión, la planeación de la misma, así como las evidencias de las evaluaciones a los niños; entregados vía Drive.

Línea Base. Se considera estable la práctica docente cuando las medias de las calificaciones presenta una variación menor al 50% del rango de puntuación (Scheeler, Congdon, y Stansbery, 2010; Scheeler, McKinnon, y Stout, 2012), con base en la REDD; durante mínimo tres sesiones cada practicante. La calificación de las practicantes es de forma escalonada, es decir, a la practicante uno y dos se le califican tres sesiones; a la tres, cuatro sesiones; y a la cuarta, cinco sesiones (ver Figura 1).

Tratamiento. La experta retroalimenta con base en los resultados obtenidos de la REDD, posterior a esto, la practicante envía su portafolio de una sesión (con las modificaciones que considere necesarias) y se vuele a calificar, se le brinda retroalimentación, etc., así sucesivamente durante mínimo tres sesiones (ver Figura 1).

Figura 1. Diseño de línea base múltiple para la calificación y retroalimentación



Estudio 2. Se pretende utilizar el enfoque cualitativo y el paradigma interpretativo, ambos según Sandin (2003); así como el estudio de caso instrumental colectivo como método de investigación acorde a Stake (1995, como se citó en Creswell, Hanson, Clark Plano, y Morales, 2007); así como la entrevista semi-estructurada y la observación como técnicas de recolección de información.

Sandin (2003) menciona que el paradigma interpretativo admite que la realidad social se construye a partir de los significados subjetivos de los individuos, es decir, permite comprender e interpretar "la realidad, los significados de las personas, percepciones, intenciones, acciones (p. 34). En cuanto al estudio de casos, éste método permite observar una situación delimitada y obtener de ella información real en momentos reales que a su vez ofrecen una visión más clara de por qué y cómo ocurren ciertos hechos y comprenden las ideas que a ellos subyacen (Cohen et al., 2007).

Las técnicas elegidas para la obtención de la información son la observación no participante y entrevista semiestructurada. Con sus respectivos instrumentos; guía de observación y guion de entrevista; ambos divididos en tres dimensiones: Autoevaluación, motivación, reflexión.

Sobre técnicas de análisis de datos, en todos los casos se realizará un análisis cualitativo del contenido, con codificación axial, la cual de acuerdo con Strauss y Corbin (2003) al relacionar las categorías con sus subcategorías se forman "explicaciones más precisas y completas de los fenómenos" (pág. 135). Cabe señalar que con la finalidad de brindar confiabilidad a la investigación, se pretende realizar la búsqueda de congruencia y pertinencia en la definición de códigos y categorías a través de jueces, además de revisar el grado de acuerdo entre las codificaciones de los analistas (Maxwell, 2009).

Al finalizar el estudio 1 se entrevistará a las practicantes para identificar los significados y experiencias de la retroalimentación de su desempeño, específicamente en su autoevaluación, motivación, reflexión y percepción. En cuanto a docente de práctica profesional será sobre el uso de la rúbrica.

Consideraciones finales

Hasta el cierre de esta ponencia se ha logrado obtener la línea base de tres practicantes con tres a cinco sesiones; ya se les brindó retroalimentación y está pendiente la entrega de su siguiente portafolio para volver a ser evaluadas. Las medias de la calificación general fueron de 1.86, 2.28 y 2.42, se espera identificar un cambio del desempeño de las practicantes, conociendo así el grado de las consecuencias de la REDD.

Referencias

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Cohen, L., Manion, L. y K. Morrison (2011). The etics of educational and social research In Research methodos in educaction. Londos / New York: Routledge.

Collie, R. y Zumbo, B. (2014). Validity Evidence in the Journal of Educational Psychology: Documenting Current Practice and a Comparison with Earlier Practice. En Zumbo, B. y Chan, E. (eds.). *Validity and Validation in Social, Behavioral, and Health Sciences, Social Indicators Research Series 54* (págs. 113-135). Suiza: Springer. DOI 10.1007/978-3-319-07794-9_7

Creswell, J. W., Hanson, W. E., Clark Plano, V. L., & Morales, A. (2007). Qualitative Research Designs: Selection and Implementation. The Counseling Psychologist, 35(2), 236–264. https://doi.org/10.1177/0011000006287390

Kazdin, A. (2000). *Modificación de la conducta y sus aplicaciones prácticas*. México: Manueal Moderno.

Kerlinger, F. y Lee, H. (2008). Investigación del comportamiento. México: McGraw-Hill.

Maxwell, J. (2009). Designing a Qualitative Study. En L. Bickman y D. Rog (Eds.). *The SAGE Handbook of Applied Social Research Methods (*214-253). Thousand Oaks, CA: SAGE Publications

Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher*, 18(2), 5–11. https:// doi.org/10.3102/0013189X018002005

Pedroza, H., y Luna, E. (2017). Desarrollo y Validación de un Instrumento para Evaluar la Práctica Docente en Educación Preescolar. *Revista Iberoamericana de Evaluación Educativa, 2017, 10*(1), 109-129. Recuperado de: https://doi.org/10.15366/riee2017.10.1.006

Pedroza, H., y García Poyato, J. (en prensa). La retroalimentación de la práctica docente, una revisión sistemática de la literatura.

Sandin, M. (2003). Perspectivas teórico epistemológicas en la investigación educativa. En *Investigación Cualitativa en Educación* (pp. 75 – 88). Barcelona: McGrawHill.

Scheeler, M. C., Congdon, M., y Stransbery, S. (2010). Providing Inmediate Feedback to Co-Teachers Through Bug-in-ear Technology: An Effective Method of Peer Coaching in Inclusion Classrooms. *Teacher Education and Special Education 33*(1), 83-96.

Scheeler, M. C., McKinnon, K., y Stout, J. (2012Effects of Immediate Feedback Delivered via Webcam and Bug-in-ear Technology on Preservice Teacher Performance. *Teacher Education and Special Education 35*(1), 77-90.

Shear, B. y Zumbo, B. (2014). What Counts as Evidence: A review of Validity Studies in Educational and Psychological Measurement. En Zumbo, B. y Chan, E. (eds.). *Validity and Validation in Social, Behavioral, and Health Sciences, Social Indicators Research Series 54* (págs. 113-135). Suiza: Springer. DOI 10.1007/978-3-319-07794-9_7

Strauss, A. y J. Corbin. (2003). *Bases de la investigación cualitativa. Técnicas y procedimientos para desarrollar la teoría fundamentada.* Colombia. Ed. Universidad de Antioquía.

Tejedor, F. (2012). Evaluación del desempeño docente. *Revista Iberoamericana de Evaluación Educativa 5*, 1e, 318-327.