Enfoques Tradicionales de Validez en la Evaluación de Clase y Credencialismo Innovador (Parte 1)

Por *Daniel Hickey*

En mi trabajo con el Laboratorio de Evaluación Participativa en la Universidad de Indiana y en mis cursos de posgrado, paso mucho tiempo ayudando a las personas a entender el concepto de *validez* en el contexto de la evaluación educativa. En este post, describo la validez como tradicionalmente se ha presentado a los educadores. Resumiré lo que un libro de texto líder ha dicho durante mucho tiempo que los educadores deben saber acerca de la validez al evaluar el aprendizaje en sus propias clases, y lo extiendo a los innovadores credencialistas que están desarrollando sistemas digitales de credencialismo, micro-credenciales y programas educativos basados en competencias. En posts posteriores, exploraré las visiones tradicionales de la "validez de cara" y la "validez consecuente". Juntos, estos posts sentarán las bases para un post final que explorará varios desarrollos nuevos en la teoría de la validez que considero importantes para estas dos comunidades.

En términos generales, la validez se refiere a la exactitud de la evidencia para apoyar reivindicaciones particulares. A diferencia de la fiabilidad, la validez *no* es una propiedad de una evaluación. Aunque muchas personas se refieren a "evaluaciones válidas", realmente no hay tal cosa. Como explicaré más adelante, la mayoría de los educadores sólo necesitan conocer algunas ideas clave cuando usan las evaluaciones en el aula, porque suelen hacer afirmaciones relativamente modestas a partir de las calificaciones de evaluación de los estudiantes. Sin embargo, mis colegas que están desarrollando sistemas innovadores de credenciales (es decir, emblemas digitales, evaluaciones de portafolio, educación basada en competencias, etc.) a menudo pretenden hacer afirmaciones más amplias sobre lo que un individuo "sabe" o "puede hacer". He trabajado diligentemente en el último año para mostrar a los "innovadores credencialistas" que es necesario un enfoque sistémico de la validez si van a prosperar sus esfuerzos. Esto incluye trabajar con:

- El Consejo Americano de Educación y sus esfuerzos para definir las Dimensiones de Calidad para las Credenciales Conectadas,
- IMS Global en sus esfuerzos para desarrollar estándares de metadatos para emblemas digitales,
- Los esfuerzos de la Fundación Mozilla para fomentar las competencias de los educadores con habilidades digitales de edad,
- Las docenas de esfuerzos para el desarrollo de emblemas que apoyamos en el proyecto de Emblemas Abiertos en Educación Superior en la Universidad de Indiana y los esfuerzos de la Alianza de Emblemas y la comunidad abierta de emblemas.

- Diseñadores y usuarios de ePortfolio, incluyendo miembros de la Asociación para el Avance del Aprendizaje Experimental y Basado en la Evidencia (AAEEBL).

Visiones Tradicionales de la Validez en la Evaluación de Clase

Durante los últimos 15 años, he estado enseñando un curso de postgrado llamado *Assessment in Schools*. Está dirigido a maestros y administradores escolares. Utilizamos el libro de Jim Popham titulado *Evaluación de Clase: Lo que los Maestros Necesitan Saber (Classroom Assessment: What Teachers Need to Know),* el libro de texto más utilizado sobre evaluación educativa en idioma inglés. Debido a que Popham reescribe el libro cada tres años, las revisiones del capítulo sobre la validez captan muy bien los aspectos más importantes de la validez para los educadores. La última (octava) edición tiene un capítulo completamente reelaborado sobre la validez. Antes de escribir sobre estos nuevos desarrollos, quiero comenzar con la visión de la validez que muchos profesionales de la medición han avanzado a lo largo de los años noventa y muchos educadores siguen aprendiendo en sus libros de texto de educación de maestros.

El libro de texto y mi curso empiezan con *objetivos* curriculares. A veces llamados *estándares educativos*, *resultados del curso* o *competencias*, estos son resultados observables que un curso se dirige a conseguir y / o que una práctica evaluativa producirá evidencia. En el caso de las enseñanzas digitales y la educación basada en competencias, éstas están representadas por listas de competencias que deben demostrarse, posiblemente independientemente de dónde y cómo se hayan adquirido esas competencias. La validez se refiere entonces a la medida en que la evidencia generada por las evaluaciones apoyan esas afirmaciones. La séptima edición de Popham, publicada en 2014, describe tres tipos de evidencia de validez:

Evidencia de validez relacionada con el contenido. Esta evidencia se refiere a "la medida en que un procedimiento de evaluación representa adecuadamente el contenido del objetivo curricular que se mide". En mi curso de evaluación, los estudiantes que trabajan como maestros concluyeron abrumadoramente que este era el tipo más relevante de evidencia de validez para ellos. Esto se debe a que están haciendo y / o usando evaluaciones en el aula que se utilizan para hacer afirmaciones sobre el dominio de los objetivos de su currículo. En mi trabajo con innovadores de acreditación, esto también resulta ser un tipo muy importante de evidencia. Como tal, lo explicaré más adelante.

Evidencia de validez relacionada con el criterio. Esta evidencia se refiere "al grado en que el desempeño en un procedimiento de evaluación predice con precisión el desempeño de un estudiante en un criterio externo". En mi curso, muchos administradores actuales o futuros concluyeron que este era el tipo más relevante de evidencia de validez para ellos. Esto se debe a que los administradores son a menudo responsables de decidir qué calificaciones en las pruebas de colocación son necesarias para demostrarlas dentro o fuera de las

clases, avanzar las calificaciones, etc. En estos casos, los "criterios externos" se refieren a si estas decisiones están asociadas con el éxito del estudiante. En mi trabajo con los innovadores de credenciales, me sorprende a menudo cuán poca atención prestan a esta evidencia, dada la naturaleza de las afirmaciones que están haciendo. Por ejemplo, los programas de educación basados en competencias a menudo asumen que alguien que ha demostrado muchas pequeñas competencias en evaluaciones de desempeño será capaz de usar esas competencias juntas en un entorno laboral real.

Evidencia de validez relacionada con la construcción. Esta evidencia se refiere "hasta qué punto la evidencia empírica confirma que existe una construcción inferida y que un procedimiento de evaluación determinado está midiendo la construcción inferida con precisión". Normalmente, sólo un puñado de estudiantes en mi clase de evaluación encuentra este tipo de evidencia más relevante. Son generalmente estudiantes de doctorado que están interesados en construcciones psicológicas como la autoeficacia o "grano". Del mismo modo, la mayoría de mis colegas de acreditación no terminan evaluando o midiendo constructos. Es una buena cosa también, ya que es bastante difícil reunir pruebas convincentes de validez relacionadas con un constructo.

En resumen, los educadores y los innovadores credencialistas a menudo hacen afirmaciones que requieren evidencia de validez relacionada con el contenido si se van a validar esas afirmaciones. Algunos administradores y diseñadores de programas a menudo reclaman evaluaciones que requieren evidencias relacionadas con un criterio. En términos generales, la evidencia de validez relacionada con la construcción se asocia con tests psicológicos y tests de rendimiento desarrollados profesionalmente.

Tipos de evidencia de validez relacionados con el contenido

El capítulo de validez de 2014 de Popham resume las maneras en que los educadores e innovadores pueden reunir evidencias, de modo que el contenido detallado en sus objetivos o competencias curriculares estén adecuadamente representados en sus evaluaciones. Para las evaluaciones de clase, la fuente más común de evidencia es simplemente el *cuidado del desarrollo*, en el cual los compañeros o los expertos examinan las evaluaciones a la luz de los objetivos curriculares o las competencias que apuntan. Para las evaluaciones de mayor participación, esto podría ampliarse a revisiones externas por grupos de expertos externos.

Ya sea que se recopile informal o formalmente, la evidencia de validez relacionada con el contenido se refiere a la alineación del contenido de la evaluación con las reclamaciones que se hacen. Si bien la alineación formal es más relevante para las pruebas de rendimiento de apuestas altas, la idea básica es la misma para las evaluaciones de clase, así como las evaluaciones de la actuación y de portafolio a menudo usadas por los innovadores de credenciales. Las cuatro "preguntas de alineación" propuestas por Noreen Webb son puntos de partida útiles:

Concordancia categórica. Esta es una indicación bastante general de si las categorías consistentes de contenido están representadas tanto en los objetivos curriculares como en la evaluación. Con las pruebas de elección múltiple, esto sería indicado por la presencia de al menos unos pocos ítems para cada objetivo curricular o estándar. Con la evaluación de la actuación y portafolio, el proceso es el mismo, pero es probable que estos sean un número menor de elementos en la rúbrica de puntuación para cada estándar. El desafío evidente aquí es que el aumento del número de objetivos o competencias curriculares requiere una evaluación más amplia.

Consistencia de profundidad de conocimiento (DOK). Indaga sobre la medida en que las demandas cognitivas de las evaluaciones son consistentes con los objetivos curriculares o competencias. Basándose en la taxonomía familiar de Bloom, Webb avanza cuatro "niveles" de profundidad, incluyendo *recordación*, *habilidad / concepto*, *pensamiento estratégico* y *pensamiento extendido*. Esta distinción es importante porque algunos currículos educativos hacen afirmaciones ambiciosas sobre el "pensamiento extendido" que va mucho más allá de lo que captan sus prácticas de evaluación. Resulta que es muy difícil crear ítems de opción múltiple o evaluaciones basadas en computadoras que provean pruebas válidas de la capacidad para resolver problemas complejos. E incluso cuando los programas desarrollan evaluaciones que capturan un conocimiento más profundo, los educadores a menudo comprometen esas evaluaciones enseñándoles directamente. (Más sobre esto en un post posterior).

Correspondencia de rango de conocimiento. Esta es una versión más específica de la primera pregunta. Se refiere a la amplitud del conocimiento representado en los objetivos curriculares en comparación con el lapso de conocimiento representado por las evaluaciones. Pero tiene en cuenta el hecho de que elementos individuales o elementos de evaluación pueden cubrir el contenido en múltiples objetivos o competencias curriculares. Esto es particularmente importante para las evaluaciones de actuación y portafolio. Se debe a que tales evaluaciones a menudo incluyen múltiples problemas extendidos o actividades que están destinados a cubrir múltiples objetivos o competencias. Esto puede llegar a estar muy desordenado muy rápidamente.

Balance de representaciones. Esto se refiere al grado en que los objetivos curriculares reciben el mismo énfasis en las evaluaciones. En otras palabras, ¿es igual la distribución del contenido en la evaluación que la distribución del contenido en los objetivos o competencias curriculares? Este es un problema relativamente sencillo con pruebas de opción múltiple que puede incluir bastantes ítems. Pero se convierte en un problema muy grande con las evaluaciones de rendimiento y portafolio que tienen menos ítems o elementos.

Uno de los puntos que espero que los lectores eliminen de este post es el desafío que presentan los formatos alternativos como el portafolio y la evaluación de la actuación para la evidencia de validez relacionada con el contenido. Aunque tanto los educadores como los estudiantes no les gustan los formatos de evaluación de opción múltiple, permiten la inclusión de muchos más ítems en una evaluación dada. Esto significa que la alineación es en su mayoría un proceso "empírico" donde el contenido y el nivel de cada elemento se

corresponde con el conjunto de objetivos o competencias. Si bien puede haber algunos desacuerdos entre los revisores, a menudo se refieren a temas específicos. Sin embargo, el proceso de alineación para la actuación y el portafolio es en gran medida un ejercicio interpretativo. Con un número menor de actividades o elementos extendidos, hay simplemente muchos más grados de libertad. Como explicaré en un post posterior, una combinación de formatos es a veces la mejor solución.

Críticas a las opiniones tradicionales de validez

A principios de la década de 1990, muchos profesionales de la medición y la evaluación comenzaron a cuestionar esta visión tradicional de la validez. Samuel Messick encabezó la acusación. Messick fue uno de los principales teóricos del *Educational Testing Service*. Su artículo de 1995 en el Psicólogo Educativo sobre "la Validez de la Evaluación Psicológica" fue uno de varios artículos más influyentes que publicó durante su vida. El resumen se abre como sigue:

La concepción tradicional de la validez la divide en tres tipos distintos y sustituibles: contenido, criterio y validez de construcción. Este punto de vista está fragmentado e incompleto, sobre todo porque no tiene en cuenta tanto la evidencia de las implicaciones de valor del significado de la puntuación como una base para la acción y las consecuencias sociales del uso de la puntuación.

Messick estaba escribiendo en parte como respuesta a la explosión de la innovación sobre evaluación a gran escala en los EE.UU. a principios de los años noventa. Me encontré en medio de estos cambios cuando empecé mi postdoc en el ETS Center for Performance Assessment, que fue creado para ayudar a informar estos esfuerzos - y en particular para ayudarles a entregar las consecuencias sociales prometidas para mejorar la enseñanza y el aprendizaje. En lugar de la tradicional visión tripartita, Messick defendió

Un nuevo concepto unificado de validez que interrelaciona estas cuestiones como aspectos fundamentales de una teoría más comprensiva de la validez del constructo que aborda tanto el significado de puntuación como los valores sociales en la interpretación de los tests y en el uso de los tests.

Samuel Messick (1931-1998)

Así, Messick trató de unificar el estudio de la validez dentro de la idea de validez de constructo, desglosada en seis "aspectos distinguibles de la validez de constructo", incluidos aspectos de *contenido, sustantivos, estructurales, generalizables, externos* y *consecuentes*. Mientras seguía el ejemplo de Popham al no presentar estas distinciones a maestros y administradores en mis clases, este marco fue central en mi trabajo de evaluación para el futuro. En particular los utilicé en un estudio de validez de las evaluaciones de actuación para la herencia introductoria que fueron centrales en mi investigación de 1996 a 2005. Como explicaré en mi próximo post, aprecié particularmente la manera en que Messick ayudó a llamar la atención sobre las consecuencias de las formas en que se utiliza la evidencia en las evaluaciones y en los tests.

(Agradezco a Daniel Hickey su autorización para traducir y publicar este post que apareció en su blog el 4 de julio de 2016. Para más información sobre su blog, vaya al final de la página principal del mío donde aparece un enlace con el suyo (Re-mediating Asseseement).

LMVA