PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2017, 54(2), 1-26

Pontificia Universidad Católica de Chile
Facultad de Educación

Examinando formas de hacer evaluación: Un acercamiento desde la teoría de respuesta del ítem para educadores de profesores

Assessing assessment literacy: An item response modeling approach for teacher educators

Brent Duckor[1], Karen Draney[2] y Mark Wilson[2]
[1]San Jose State University
[2]University of California, Berkeley

Resumen

El presente estudio examina como distinguir de manera significativa y sistemática los niveles de evaluación de conocimiento en aula entre profesores practicantes, utilizando un marco de progresión del aprendizaje. Guiados por principios nacionalmente reconocidos sobre el desarrollo y adquisición de evaluación en aula (CAL por sus siglas en inglés), el instrumento fue usado para identificar las diferencias cualitativas y cuantitativas sobre evaluación de aula de educadores principiantes. El resultado de rendimiento en la escala presenta nuevas direcciones para modelar el entrenamiento de los profesores practicantes y sus progresiones en este complejo ámbito. Los descubrimientos preliminares resultan relevantes para educadores que utilicen evaluaciones y estén interesados en usos diagnósticos y formativos del instrumento presentado para evaluar el crecimiento de los profesores practicantes.

Keywords: progresiones de aprendizaje, teoría de evaluación, educación de profesores en formación, teoría de respuesta del ítem, validación, fiabilidad

Correspondencia a:
Brent Duckor, Associate Professor
Department of Secondary Education SH 436
Lurie College of Education
San Jose State University
Email: Brent.Duckor@sjsu.edu

ISSN:0719-0409 DDI:203.262, Santiago, Chile
doi: 10.7764/PEL.54.2.2017.5

Abstract

The study articulates how to meaningfully and consistently distinguish between levels of classroom assessment knowledge among pre-service teachers using a learning progressions framework. Guided by nationally recognized principles of development and acquisition of assessment literacy, the Classroom Assessment Literacy (CAL) instrument used to identify qualitative and quantitative differences in beginning classroom assessors' proficiency estimates on a scale yields insight into new directions for modeling teacher learning progressions in this complex domain. Preliminary findings are relevant to assessment educators who are interested in diagnostic and formative uses of the CAL scales for evaluating pre-service teacher growth.

Keywords: assessment literacy; learning progressions; pre-service teacher education; item response theory; validity; reliability

En vísperas de la promulgación de la ley No Child Left Behind [Que ningún niño se quede atrás] (2001), los expertos en evaluación en los Estados Unidos hicieron un llamado para aumentar la "alfabetización evaluativa" en profesores, funcionarios y administradores en el campo de la educación preescolar, primaria y secundaria. Por más de dos décadas, los investigadores se han interesado en el tema de la alfabetización evaluativa y su importancia para profesores, estudiantes y padres (Stiggins, 2001, 2002; Marzano, Pickering, & Pollock, 2001; Popham, 2004; Darling-Hammond, 2006; Gotch & French, 2014). Mientras estos expertos presionaron para que el tema se discutiera con mayor profundidad, los legisladores a nivel nacional y estatal en los Estados Unidos trataron de usar un lenguaje claro y sólido con respecto al papel de la evaluación en sus documentos sobre estándares profesionales (CCSSO, 2010; NBPTS, 2012; NRC, 2010). Al señalar tanto su importancia para la práctica profesional como su influencia más general conducente a una reforma educativa desde el nivel preescolar al secundario, el término "alfabetización evaluativa" se convirtió en un estandarte.

Con el surgimiento del movimiento de los Estándares Centrales Comunes [Common Core Standards] y la visión de políticas de evaluación a gran escala articulada por grupos como el Consorcio para una evaluación más inteligente [Smarter Assessment Consortium] en los Estados Unidos, el hincapié en los conocimientos sobre evaluación de los profesores en formación se ha acentuado en los últimos años. En California, por ejemplo, los estándares profesionales para profesores entregan múltiples metas de aprendizaje para profesores en formación, las cuales se centran en prácticas de evaluación en aula (CCTC, 2012). La versión corregida del documento Teaching Performance Expectations [Expectativas de desempeño para profesores] (CCTC, 2016), por ejemplo, describe cuidadosamente el conjunto de conocimientos, habilidades y capacidades que se requieren de acuerdo a las directrices del TPE 5 (Assessing Student Learning; Evaluación del aprendizaje de los estudiantes) para que un candidato a recibir el título de profesor pueda obtener una licencia para enseñar en California.

Aunque los documentos que definen los estándares profesionales y los organismos que conceden licencias para profesores han sido centrales para establecer en qué consiste la "alfabetización evaluativa", debido a la Carrera a la cima [Race to the Top] se han dejado de lado otros enfoques y perspectivas con respecto a la comprensión y el uso de la evaluación en el aula por parte de los profesores. Una

consecuencia de la era de la rendición de cuentas (accountability) y las pruebas de "alto impacto", impulsada estatalmente en los Estados Unidos, fue que se redujo la importancia del papel del profesor como aprendiz y que los ejemplos de experticia del profesor en el ámbito de la evaluación en el aula quedaron relegados a reformas educacionales pasadas (Duckor & Perlstein, 2014).

En este artículo, exploramos la idea de las progresiones de aprendizaje de los profesores, centrándonos particularmente en cómo los profesores principiantes van superándose hasta comprender y usar de forma más sofisticada las prácticas de evaluación en un programa de formación pedagógica (ver, por ejemplo, Shavelson, Moss, Wilson, Duckor, Baron, & Wilmot, 2010; Duckor, 2017). El concepto del profesor como aprendiz exige que definamos *continuos* hipotéticos de práctica profesional para profesores en formación. Sobre la base de Estándares pedagógicos reconocidos nacionalmente y el informe del Consejo Nacional de Investigación [National Research Council] (2001a) denominado *Saber qué es lo que saben los estudiantes: la ciencia y el diseño de la evaluación educativa* [Knowing what students know: The science and design of educational assessment], investigamos empíricamente cómo los profesores principiantes pueden abordar y aprender sobre la lógica de la evaluación dentro de un contexto de aprendizaje particular: los programas de formación para profesores en los Estados Unidos, específicamente en California.

Como formadores de profesores y especialistas en medición educacional, sostenemos que parte del aprendizaje de la "lógica de la evaluación en el aula" involucra desarrollar modelos y esquemas mentales más sofisticados. Éste también implica oportunidades para usar herramientas mentales y lingüísticas y aplicar conocimientos sobre evaluación en distintos contextos y áreas de la práctica profesional (Vygotsky, 1978). La investigación sobre progresiones de aprendizaje en profesores podría ayudar a describir y evaluar estas trayectorias cognitivas en principiantes, especialmente en la intersección entre las experiencias en el trabajo clínico de campo y las experiencias de aula en la universidad. Como sugiere el título del artículo, nos interesa la emergencia y desarrollo de la alfabetización evaluativa en el aula en profesores en formación que buscan aumentar sus conocimientos sobre evaluación antes de desempeñarse como profesionales en su área.

Las preguntas globales que motivan este estudio son: ¿Están preparados los profesores en formación para adoptar y promover la nueva alfabetización educativa visualizada por los expertos? ¿Qué tipos de modelos mentales (por ejemplo, p-prims y concepciones erróneas) tienen los profesores en formación acerca del tema de la evaluación educativa, por ejemplo, con respecto a las calificaciones o las puntuaciones? ¿De qué manera podrían entender los formadores de profesores el constructo de la "alfabetización evaluativa", para así entregar a los profesores en formación maneras de pensar más potentes sobre la evaluación en el aula hoy? En California, los más recientes modelos estatales de rendición de cuentas (accountability) que promueven el mejoramiento continuo, sumados a evaluaciones de desempeño para profesores alineadas con prácticas de evaluación más profundas a nivel de aula, vaticinan una comprensión nueva y más robusta de la alfabetización evaluativa.

Trasfondo y contexto

Los expertos en evaluación, medición y administración de tests en el ámbito educacional coinciden en que la *cognición*, la *observación* y la *interpretación* son esenciales para entender la evaluación tanto en el aula como a gran escala (NRC, 2001a). Cada uno de estos componentes forma parte de la lógica subyacente a cualquier sistema de evaluación: juntos, estos vértices ("temas") entregan evidencia que apoya los esfuerzos de validación, por ejemplo, para fomentar el uso justo y apropiado de los datos.

Sobre la base del modelo mental del "Triángulo de la evaluación", presentado por los expertos del NRC, exploramos la evidencia con respecto a las progresiones de aprendizaje de los profesores en formación empleando componentes similares y la lógica desarrollada originalmente por estos especialistas en evaluación. En la Figura 1 se muestra una versión modificada del Triángulo de la evaluación del NRC:

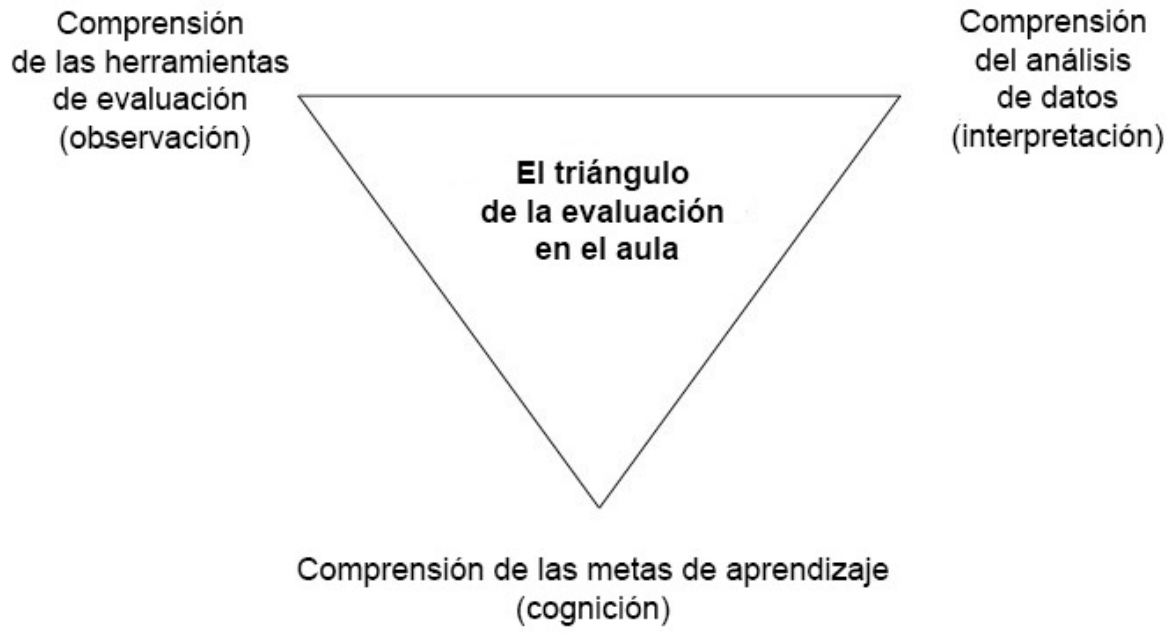


Figura 1. Triángulo de la evaluación en el aula (adaptado de NRC, 2001).

Nuestra definición de alfabetización evaluativa se inspira en gran medida en la lógica del diseño y el uso de la evaluación mostrada en la Figura 1, con un énfasis particular en cómo los profesores pueden y deben usar sus propias evaluaciones en el aula. Cada vértice ("Metas de aprendizaje", "Herramientas de evaluación" e "Interpretación de datos") representa la lógica de la evaluación en el aula con respecto a tres temas centrales, empleando términos más accesibles para los profesores en formación.

Nuestra hipótesis es que el conocimiento sobre el Triángulo de evaluación en el aula (TEA; Classroom Assessment Triangle, CAT) tiene al menos tres dominios separados, cada uno con su respectiva progresión de aprendizaje para poblaciones de profesores en formación.

La primera progresión de aprendizaje, "Comprensión de las metas de aprendizaje", describe la comprensión que tiene un profesor sobre la cognición de los estudiantes como parte de la evaluación en el aula. Ésta define diferencias en la sofisticación de los profesores en formación con respecto al reconocimiento de exigencias cognitivas, habilidades y niveles de comprensión en situaciones de evaluación en el aula. Por ejemplo, una característica de esta progresión es la capacidad del profesor para conectar metas de aprendizaje con la evaluación de habilidades de pensamiento de orden superior e inferior.

La segunda progresión de aprendizaje, "Comprensión de las herramientas de evaluación", describe la comprensión de los profesores en formación sobre las posibilidades y limitaciones de distintas herramientas y estrategias de evaluación. Ésta define un rango de comprensiones con respecto a los formatos, modalidades y características de lo que los expertos entienden como diseño de ítems. La habilidad del profesor en formación para diseñar y mejorar estas herramientas de evaluación, incluyendo su capacidad de prever procedimientos alternativos de puntuación o de generación de sentido, es un elemento clave en esta progresión.

La tercera progresión de aprendizaje, "Comprensión de la interpretación de datos", describe la comprensión que tienen los profesores en formación sobre la calidad de los datos de aula, incluyendo su capacidad de hacer inferencias y sacar conclusiones válidas sobre la base de puntuaciones. El conocimiento y capacidad del profesor en formación para recolectar y evaluar distintos tipos de evidencia sobre validez y fiabilidad que le permitan defender (o criticar) el uso de un instrumento constituyen una característica esencial de esta progresión.

Para operacionalizar estas hipotéticas progresiones de aprendizaje, examinamos tres preguntas de investigación: 1) ¿Existen niveles distintos de comprensión de los temas de una versión modificada del Triángulo de Evaluación NRC? 2) ¿De ser así, cómo se puede distinguir significativa y sistemáticamente entre estos niveles de comprensión de los profesores en una escala psicométricamente sólida? y 3) ¿Existe alguna evidencia de fiabilidad y validez de las puntuaciones de la escala que permita justificar usos diagnósticos y formativos para educadores de profesores?[1]

Así, un objetivo central del presente estudio sobre progresiones de aprendizaje consiste en medir a un grupo de profesores y calibrar ítems en tres áreas temáticas que cubren los dominios principales de un marco nacionalmente reconocido para examinar el saber evaluativo. Empleando el modelo Rasch, se ajustaron los datos generados por un instrumento pre-post test para examinar el continuo propuesto de niveles de alfabetización evaluativa con una muestra de profesores en formación. En lugar de adoptar la diferenciación entre experto-principiante que comúnmente se presenta en la literatura sobre evaluación, se definió el espacio del constructo de *Alfabetización sobre evaluación en el aula* (AEA; CAL, Classroom Assessment Literacy) como un conjunto potencial de progresiones de aprendizaje que pueden estimarse con una escala usando métodos cuantitativos (por ejemplo, véase Duckor, Draney, & Wilson, 2009). También se empleo el enfoque de construcción de medidas (Wilson, 2005) para investigar el constructo mismo de "alfabetización evaluativa". Nuestro objetivo como formadores de profesores y expertos en psicometría es determinar hasta qué punto existe o no una variación significativa entre profesores individuales dentro de un continuo claramente definido de desempeños basados en tareas. En la sección siguiente se presenta una definición del constructo de "alfabetización evaluativa" que puede someterse a investigación empírica y psicométrica dentro del currículo de un programa de formación de profesores en una universidad estatal de tamaño grande.

Marco analítico y metodología

La tarea de definir en qué consiste el conocimiento experto sobre evaluación en el aula en términos de "alfabetización" representa un desafío y probablemente cause controversia.[2] Sin embargo, una revisión de los Estándares para profesores tanto estatales como nacionales revela ciertas características comunes de un constructo psicológico que puede considerarse, provisionalmente, como un tipo de alfabetización evaluativa para profesores (CCTC, 2012, 2016; NBPTS, 2012). Los *Estándares de competencia pedagógica en la evaluación educativa de los estudiantes* [Standards for Teacher Competence in the Educational Assessment of Students] (AFT, NCME, & NEA, 1990) definen evaluación como

"el proceso de obtener información que se usa para tomar decisiones educativas sobre los estudiantes, dar retroalimentación al estudiante sobre sus avances, fortalezas y debilidades, hacer juicios sobre la efectividad de la enseñanza y la adecuación del currículo e informar políticas en el ámbito de la educación". Los siete Estándares de 1990 entregan criterios sobre la competencia de los profesores con respecto a los múltiples componentes de esta amplia definición de evaluación.[3]

Si bien este estudio emplea los Estándares para definir la alfabetización sobre evaluación en el aula, además se busca describir el contenido y estructura de las variables latentes mediante un enfoque particular de modelamiento de constructos (Wilson, 2005), el cual hace hincapié en el diseño centrado en evidencias (Mislevy, Almond, & Lukas, 2003) y en los avances en los campos de la teoría de respuesta al ítem y la medición. En este estudio de teoría de respuesta al ítem tipo Rasch, nuestra meta primaria es construir una medida de los profesores en términos de su "competencia" latente y calibrar los ítems de acuerdo a su "dificultad" dentro de una escala técnicamente sólida. También nos interesa validar las puntuaciones derivadas del instrumento CAL (Classroom Assessment Literacy; Alfabetización sobre evaluación en el aula) sobre la base de una investigación de la evidencia sobre su validez y fiabilidad. Los usos apropiados del instrumento CAL dependen en gran medida de la estabilidad y significación de las inferencias justificables de acuerdo a los Estándares (AERA, APA, NCME, 1999, 2014).

Para conceptualizar nuestras ideas iniciales sobre la estructura del aprendizaje en la variable CAL, empleamos la taxonomía denominada Estructura del resultado de aprendizaje [Structure of the Learning Outcome, SOLO]. La taxonomía SOLO (Biggs & Collis, 1982) es un marco teórico general que puede usarse para construir una estrategia de puntuación o codificación dirigida a hacer emerger el nivel de sofisticación cognitiva de un sujeto (en este caso, el evaluador del aula) mediante una combinación de ítems de opción fija y tareas de desempeño escrito. En el presente estudio, se usa el enfoque SOLO para pasar de una jerarquía de etapas ontológicas fijas (experto o principiante) a una jerarquía de categorías de resultado observables (de discordante a integrativa) para evaluar a profesores cuyo dominio en un campo particular de la práctica está desarrollándose a través del tiempo.

En el marco CAL, se espera que los evaluadores de aula (categoría que no solamente incluye a los profesores en formación) se nutran de al menos tres temas de conocimiento para demostrar su comprensión de la evaluación en el aula. Aunque sospechamos que algunas de las competencias involucradas en cada tema pueden estar fuertemente relacionadas entre ellas, igualmente buscamos distinguir cuidadosamente los temas en la fase de definición del constructo. Por lo tanto, se desarrolló un total de cuatro mapas de constructo para representar a cada uno de los temas dentro del marco CAL.

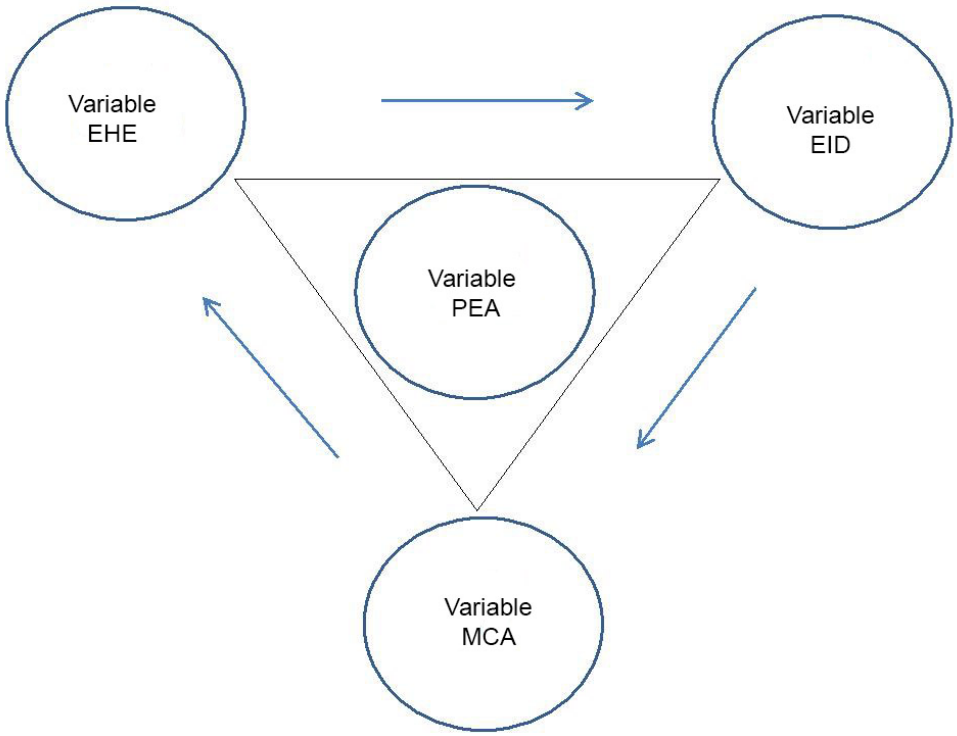


Figura 2. Relaciones entre temas de los dominios de las variables CAL (Classroom Assessment Literacy; Alfabetización sobre evaluación en el aula).

En el primer tema se encuentra el mapa de constructo Comprensión de las metas de cognición y aprendizaje (o "variable MCA"), el cual se centra en los tipos y calidad de las representaciones que usa el evaluador para definir una meta de evaluación. Esta variable también se refiere a la comprensión que tienen los profesores sobre las propiedades, cualidades y limitaciones de procedimientos de mapeo específicos. La variable MCA representa la habilidad de los evaluadores de aula para diseñar y evaluar las metas de aprendizaje de los estudiantes considerando, entre otros criterios, las oportunidades de observar pensamiento de orden superior e inferior que son capaces de generar.

El segundo tema es el mapa de constructo Comprensión de las estrategias y herramientas de evaluación (o "variable EHE"). Esta variable se centra en el conocimiento que tiene el evaluador sobre los formatos y usos tradicionales de los ítems, además de las reglas generales para construir "buenos" ítems. Esta variable también comprende la visión más sofisticada de los ítems como muestras de un universo que podría (o no) tener una relación plausible con el pensamiento de los estudiantes, incluyendo sus ideas incorrectas y sus errores de comprensión, pero sin limitarse a ellos. La variable EHE representa la habilidad del evaluador de aula para diseñar y examinar actividades de evaluación (por ejemplo, preguntas, tareas, ítems y tests) que se alineen con las tareas de aprendizaje definidas, por ejemplo, por la variable MCA.

El tercer tema es el mapa de constructo Comprensión de la evidencia y la interpretación de datos (o "variable EID"). Incluye el conocimiento del evaluador de aula sobre las propiedades de las estrategias de puntuación y el uso que hace de ellas, de acuerdo a su propósito, contexto y uso. Esta variable se refiere a una noción más sofisticada de las rúbricas, claves de respuesta y otras herramientas

de calificación como métodos de generar resultados que pueden usarse con fines sumativos, formativos y diagnósticos. Este enfoque centrado en las interpretaciones y usos de los datos de los estudiantes implica un interés mayor en la validación, incluyendo la fiabilidad de los resultados. Si las rúbricas, por ejemplo, no están alineadas con los estándares de contenidos, niveles taxonómicos o resultados cognitivos, las inferencias que se hagan sobre el aprendizaje de los estudiantes serán imprecisas. De este modo, la variable EID representa la capacidad del evaluador de aula para diseñar, evaluar y modificar estrategias de puntuación que se alineen con los elementos de las variables MCA (metas de aprendizaje de los estudiantes) y EHE (estrategias y herramientas de evaluación), respectivamente.[4]

Es importante tener en cuenta que cada mapa de constructo se caracteriza por la variación (o un "continuo") de niveles de desempeño con respecto a los evaluadores de aula ("personas") y a las respuestas a tareas ("ítems"). El mapa de constructo articula la estructura esperada de los resultados de acuerdo a la dificultad de los ítems/tareas (Wilson, 2005). Entrega un esquema de codificación generalizado que finalmente será mapeado según guías específicas de puntuación. Los evaluadores de aula pueden o no adecuarse a estas expectativas (de constructo); por lo tanto, el mapa está sujeto a modificaciones basadas en hallazgos empíricos. La Figura 3 entrega un ejemplo del mapa de constructo desarrollado para el dominio temático MCA.

Figura 3. Mapa de constructo de Comprensión de las metas de cognición y aprendizaje (Variable MCA).

Como se observa en la Figura 3, en el extremo inferior del mapa de constructo MCA proponemos la existencia de evaluadores de aula principiantes (es decir, los que se hallan en el nivel de "pre-evaluación"), quienes aún no tienen la experiencia para desarrollar, analizar y modificar diferentes aspectos del elemento de la cognición de los estudiantes dentro del triángulo de la evaluación. Estos individuos tienden a ignorar o a no prestar atención a ninguna noción relacionada con las bases socioculturales de la evaluación en el aula (por ejemplo, véase Shepard, 2000). Pueden tener creencias fijas sobre el carácter de los resultados de aprendizaje y la inteligencia de los estudiantes (por ejemplo, véase Dweck, 2010), lo que puede contribuir a sus dificultades para conceptualizar un continuo de desempeño estudiantil. Su conocimiento sobre prácticas, principios y "movimientos" de evaluación es inconexo e incompleto (por ejemplo, véase Berliner, 1988; Ball & Cohen, 1999; Duckor, 2014, Duckor & Holmberg, 2017). Los evaluadores de aula que se encuentran en este nivel suelen preguntarse "¿Qué tiene que ver la evaluación con la enseñanza"? (Popham, 2007b), podrían afirmar "Yo evalúo poniendo notas" (Guskey, 2006) o tal vez crean que "una nota es una nota" (Braun & Mislevy, 2005).

En el nivel superior del mapa de constructo MCA, proponemos la existencia de un grupo de expertos en evaluación en el aula (es decir, los que se hallan en el nivel "integrativo"), quienes son capaces de identificar y usar varios modelos mentales para representar aspectos cognitivos, observacionales e interpretativos de la evaluación. Estos individuos pueden usar de forma flexible y adaptar los elementos del *Triángulo de evaluación en el aula* (o modelos mentales similares) al tiempo que reconocen las potenciales cualidades y limitaciones de una práctica de evaluación específica. Estos individuos entienden que los propósitos y contextos condicionan cómo se diseñan y se modifican las herramientas, procedimientos y estrategias de evaluación para apoyar a los estudiantes.[5] Los evaluadores de aula expertos también tienen en cuenta la evidencia a favor y en contra de sus juicios sobre los avances del aprendiz de acuerdo a conceptos como estandarización ("acceso"), validez ("significación") y fiabilidad ("consistencia"). Es decir, al examinar los datos de puntuación de los estudiantes para hacer un juicio, estos individuos reconocen el carácter provisional de las inferencias y la necesidad de tener evidencias para apoyar las evaluaciones sobre el aprendizaje de los estudiantes.

Aunque estamos bastante confiados de haber identificado los extremos en nuestro mapa de la variable MCA, también nos interesan los niveles intermedios ubicados entre los "expertos" y "principiantes" en la evaluación en el aula, especialmente dado nuestro rol como formadores de profesores. Hemos observado que los encuentros tempranos con la identificación de metas de aprendizaje en un plan de una unidad de enseñanza suelen dejar en un estado "discordante" (es decir, el nivel 2 de la Figura 3) a quienes se hallan estudiando la evaluación en el aula. Por ejemplo, estas personas son capaces de describir y nombrar múltiples metas, estándares y objetivos que desean evaluar en un plan de una unidad de enseñanza; sin embargo, tienen dificultades para concentrarse en definir metas de aprendizaje individuales para una clase o conjunto de clases. Algunos investigadores (Heritage, Kim, & Vendlinski, 2008) han subrayado la importancia de "la claridad que tienen los profesores sobre lo que viene antes o después de una meta de aprendizaje específica." (p.4). En otras palabras, sin un marco mental –como por ejemplo una "progresión de aprendizaje" o una idea sobre las "facetas de la comprensión"– puede ser difícil para estos profesores ensamblar todas las "partes móviles" que existen. En este nivel, observamos que las habilidades cognitivas que tienen los profesores para crear mapas que revelen patrones en el pensamiento de los estudiantes (por ejemplo, identificar "ideas erróneas" comunes en una cierta área de contenidos) no están bien definidas, no son suficientemente exhaustivas ni están ordenadas de maneras particularmente significativas.[6]

El siguiente nivel en la progresión de la alfabetización evaluativa para la variable MCA lo denominamos "emergente" (Fig. 3), en parte porque se refiere a individuos que están emergiendo como

usuarios competentes de una estrategia específica de mapeo cognitivo como los mapas conceptuales o de estudiantes. Estos evaluadores de aula están empezando a enfrentar la complejidad cognitiva en los contenidos de su asignatura, en parte, tratando de descomponer o agrupar fenómenos en múltiples "sub-metas" dentro de sus estrategias para planificar clases y unidades. Las prácticas de evaluación en el aula en este nivel buscan alinear los controles, pruebas y tareas de desempeño en el aula con Estándares o tipos de conocimiento, lo que entrega una justificación más sustantiva para asignar calificaciones o puntuaciones. En este nivel de competencia, los individuos normalmente logran identificar aspectos que podrían llevar a error insertos en las metas de aprendizaje, comúnmente definidas como "idea central", "foco central", u "objetivo", entre otras denominaciones.

Una respuesta emergente típica al analizar un mapa de planificación de evaluaciones en la asignatura de Inglés, por ejemplo, comienza por separar las "habilidades" de lectura, escritura y fluidez oral. El profesor principiante reconoce estas metas de aprendizaje como elementos separados pero puede seguir teniendo problemas para entender el ensayo persuasivo y sus múltiples elementos, como la voz, la estructura, la tesis y/o las convenciones de escritura. Con demasiada frecuencia, vemos que los profesores en formación en este nivel intentan aplicar reflexivamente la taxonomía de Bloom y la "hacen coincidir" con el problema de representar y por lo tanto evaluar las metas y habilidades de aprendizaje de los estudiantes; asimismo, en algunos casos recurren a enunciados del tipo "los estudiantes serán capaces de" con respecto al papel de una herramienta o actividad de evaluación discreta dentro de una planificación de clase.

En la zona intermedia de la progresión del aprendizaje del profesor, proponemos la existencia de un nivel "multiestructural" de competencia en la variable MCA (Fig. 3). En este nivel, cada profesor tiene la capacidad de explicar por qué y cómo sus prácticas de evaluación se relacionan con la cognición, las teorías socioculturales sobre el aprendizaje y el aprendiz (por ejemplo, véase Shepard, 2000). Estos evaluadores muestran un conocimiento esquemático sobre múltiples prácticas de evaluación en el aula, el cual coordinan al tratar de alinear ítems, estrategias de calificación y metas de aprendizaje para fortalecer la cadena de inferencia (de la puntuación al juicio). Se centran mucho menos en las prácticas de calificación y privilegian llevar un registro de los patrones de respuesta de los estudiantes, estableciendo conexiones con patrones de aprendizaje y enseñando de acuerdo a las exigencias de contenido específicas el tema de la unidad.

Es probable que los profesores que se hallan en este nivel intermedio tengan suficientes conocimientos como para definir metas de corto plazo que cubran porciones manejables de instrucción y coordinar estrategias de evaluación formativa, al tiempo que identifican el objetivo de cada clase dentro de una trayectoria de instrucción que apoye el aprendizaje de los estudiantes a través del tiempo (Alonzo & Gearhart, 2006). A veces denominados "profesores expertos", estos individuos también son expertos en evaluación en el aula que dominan en algún grado las ideas relativas a validez, fiabilidad y sesgo del ítem. Muchos de estos profesores han participado en sesiones de calibración de exámenes de graduación en sus escuelas, han dirigido debates a nivel de escuela sobre calificación basada en estándares o han actuado como representantes frente a las autoridades del distrito en el desarrollo de ítems, tareas y pruebas comparativas (por ejemplo, véase Darling-Hammond, Ancess, & Falk, 1995).

Luego de haber definido la estructura del marco CAL propuesto, en parte mediante la presentación de un ejemplo del contenido de dominio para el tema MCA, pasamos a examinar la evidencia que apoya o cuestiona la teoría subyacente al constructo de "competencia" en el marco NRC modificado (2001a). Nuestro interés primario en estos niveles de desarrollo de la competencia es lograr una mejor comprensión de las diferencias de desempeño de los profesores para así mejorar los resultados de

aprendizaje para “principiantes” y “expertos”; además, lo que podría ser aún más importante, buscamos mejorar los aprendizajes de quienes se encuentran entre esas dos categorías de investigación tradicionales (por ejemplo, véase Borko & Livingston, 1989; Putnam & Borko, 2000; Feiman-Nemser, 2001).

Se realizó una codificación provisional de estos niveles de desempeño (pre-evaluación, discordante, emergente, multiestructural e integrativo) para refinar la teoría del constructo, la que representa un continuo de conocimientos y competencia con respecto a la evaluación. Sobre la base de principios nacionalmente reconocidos de diseño y práctica en el campo de la evaluación, proponemos que existen diferencias cualitativamente importantes tanto en la comprensión como en la práctica de la evaluación en el aula. Enumeramos la lógica de la evaluación en el aula según el NRC en lugar de hacer un inventario de la amplia gama de habilidades que comúnmente se bosquejan en los estándares profesionales para profesores.

La sección siguiente se refiere a los métodos cuantitativos y a las fuentes de datos usadas para investigar nuestras hipótesis sobre la estructura y funcionamiento de estas variables en una muestra de profesores en formación en una universidad diversa y de tamaño grande que prepara a candidatos a enseñar una única asignatura en el Estado de California.

Métodos y fuentes de datos

Descripción de los participantes

Muestra. Para este estudio, se obtuvo una muestra de 72 individuos inscritos en tres secciones de distintos cursos. Cada sección estaba compuesta de profesores en formación pertenecientes a un programa de post-licenciatura ofrecido a candidatos a enseñar una única asignatura en una universidad de tamaño grande ubicada en el Estado de California. El curso es impartido por la Escuela de Educación y se dicta en paralelo a un programa de práctica profesional de Fase II/III en distintas escuelas secundarias del norte de California. La Tabla 1 muestra las características demográficas de los participantes según conteo y porcentaje de frecuencia.

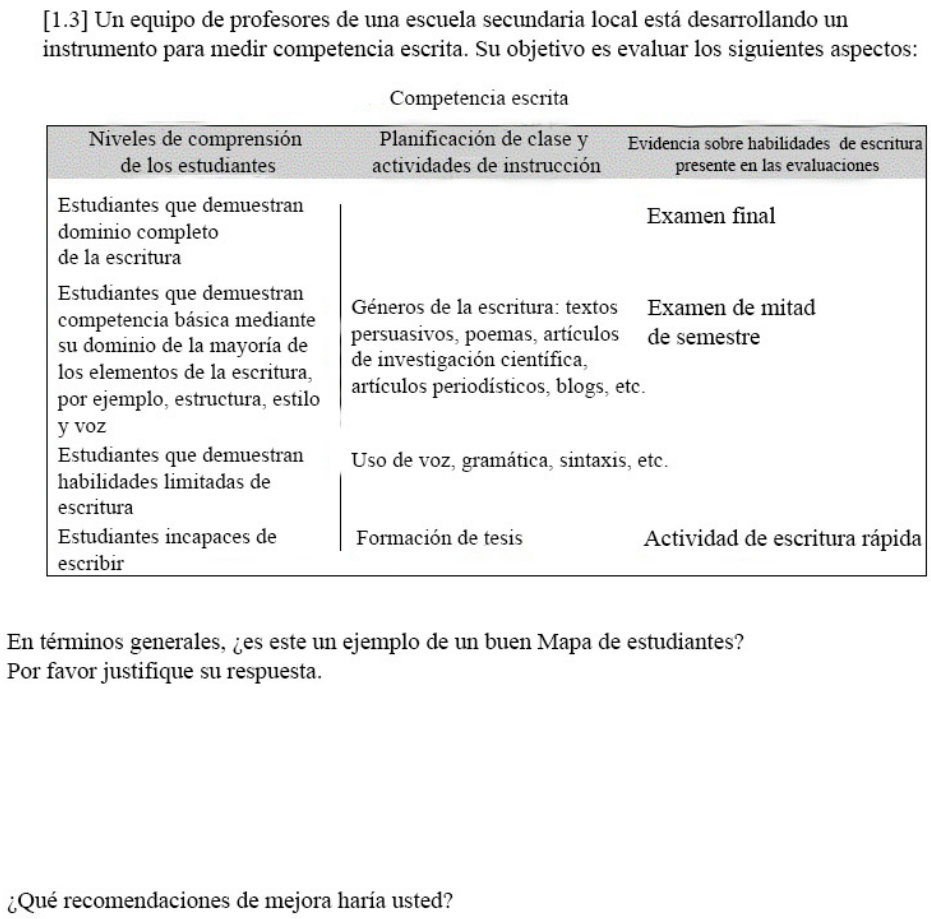
Tabla 1
Características demográficas de la muestra seleccionada (n=72)

Característica	N	Porcentaje de frecuencia
Mujer	34	47,2%
Caucásica	47	65,3%
Más de 40 años	14	19,4%
Asignatura		
Matemática	11	15,3%
Ciencia	11	15,3%
Otra	50	69,4%

El método principal de reclutamiento fue la comunicación oral. La muestra se obtuvo con consentimiento informado aprobado por el Comité de Revisión Institucional y los datos se recolectaron como parte de las actividades académicas normales. Además, se incluyeron cuatro ítems de salida de la entrevista en el instrumento.

Instrumentos

Diseño de los ítems. El instrumento CAL es un test de competencia pre-post diseñado para medir la comprensión y el uso del marco NRC (2001a), con un énfasis particular en los cuatro dominios temáticos relacionados con el Triángulo de la evaluación. El test está compuesto por 55 ítems: 13 preguntas de respuesta construida y 42 de elección fija. Cada ítem apunta a un único dominio del marco CAL y está diseñado para cubrir distintas partes de un mapa de constructo específico dentro del marco CAL. En la Figura 4 se muestra un ejemplo de un ítem de respuesta construida perteneciente al dominio MCA.



[1.3] Un equipo de profesores de una escuela secundaria local está desarrollando un instrumento para medir competencia escrita. Su objetivo es evaluar los siguientes aspectos:

Competencia escrita

Niveles de comprensión de los estudiantes	Planificación de clase y actividades de instrucción	Evidencia sobre habilidades de escritura presente en las evaluaciones
Estudiantes que demuestran dominio completo de la escritura		Examen final
Estudiantes que demuestran competencia básica mediante su dominio de la mayoría de los elementos de la escritura, por ejemplo, estructura, estilo y voz	Géneros de la escritura: textos persuasivos, poemas, artículos de investigación científica, artículos periodísticos, blogs, etc.	Examen de mitad de semestre
Estudiantes que demuestran habilidades limitadas de escritura	Uso de voz, gramática, sintaxis, etc.	
Estudiantes incapaces de escribir	Formación de tesis	Actividad de escritura rápida

En términos generales, ¿es este un ejemplo de un buen Mapa de estudiantes? Por favor justifique su respuesta.

¿Qué recomendaciones de mejora haría usted?

Figura 4. Ítem de respuesta construida perteneciente al Dominio MCA del Instrumento CAL

Este ítem es un ejemplo típico del formato de respuesta construida que se usa en el Instrumento CAL. Fue diseñado para explorar cómo los participantes comprendían la especificidad, la direccionalidad y el ordenamiento de los resultados cognitivos de los estudiantes considerando su relación con la tarea de mapeo de los aprendices. El ítem presenta una situación escrita junto con una representación de un Mapa de estudiantes pobremente diseñado. Hay dos consignas abiertas que requieren una respuesta corta. De modo similar al formato de tareas de desempeño del PACT (Performance Assessment for California Teachers [Evaluación de Rendimiento para Profesores de California]), se espera que los participantes entreguen una explicación escrita que justifique el uso de sus propias herramientas de aula.

Procedimiento de puntuación. Se empleó una estrategia de puntuación politómica para este conjunto de datos. Se usaron guías de asignación de puntuaciones para codificar respuestas para los ítems de respuesta construida y selección fija. Cada guía de puntuación se diseñó para que se alineara con los mapas de constructo CAL, los que fueron considerados como el "espacio de resultados" generalizado (Wilson, 2005). Iterando desde los requerimientos de codificación generales a los específicos, se llegó a lo que denominamos guías de asignación de puntuaciones. En la Figura 5 se muestra un ejemplo para el dominio temático MCA.

Dominio Metas de cognición y aprendizaje	Ítem (1.3)	Descripciones de ejemplares Representando los resultados de aprendizaje con un Mapa de estudiantes
Integrativo	4	Comprende las propiedades de una meta de aprendizaje bien definida y única (constructo) □ Podría expresar preocupaciones con respecto a la validez de múltiples metas (dimensionalidad)
Multiestructural	3	Entre otras cosas, identifica problemas con respecto a la definición de la meta de aprendizaje (constructo) □ Entrega consejos relevantes; por ejemplo, usa la Taxonomía de Bloom para definir metas O se centra en los avances relativos a una idea central amplia O crea "mapas múltiples" para capturar mejor la trayectoria
Emergente	2	Reconoce al menos una característica convencional de superficie, por ejemplo brechas, niveles o tipos de evidencia □ Entrega consejos genéricos; por ejemplo "faltan descripciones específicas", "añadir más niveles" o "faltan ejemplos del trabajo de los estudiantes" □ Podría *presuponer* que es necesario aplicar una taxonomía específica, como por ejemplo los niveles de Bloom
Discordante	1	Afirma que "Se ve bien" O entrega consejos vagos y que llevan a error
Pre-evaluación	0	No responde o no se refiere al tema consultado

Figura 5. Guía de puntuación alineada con la Variable CAL generalizada

Estas guías de asignación de puntuaciones fueron diseñadas para que se alinearan con el espacio de resultados generalizado, preservando así la estructura global de la variable; en otras palabras, las categorías usadas para describir niveles de competencia y dificultad de los ítems se mantienen constantes en todas las guías. La razón principal para utilizar estas guías "ejemplares" fue que sirven de apoyo para los protocolos de puntuación de los observadores. Todos los ítems de respuesta construida para los tests pre y post del presente estudio fueron puntuados de manera ciega por el autor principal (es decir, se eliminaron los datos de identificación personal en todas las respuestas) para reducir los potenciales sesgos intraobservador.[7]

Además de las guías de puntuación generadas para cada área temática, también se empleó una estrategia de diseño de Opción Múltiple Ordenada para los ítems de selección fija (por ejemplo, véase Briggs, Alonzo, Schwab, & Wilson, 2006). Esto posibilitó justificar la asignación de puntuaciones parciales y mejoró nuestra comprensión de las razones que explican la estructura de los datos. Estas guías de asignación de puntuaciones también permiten mayor flexibilidad en la especificación de modelos de medición, por ejemplo, al explorar la codificación politómica o dicotómica de los datos generados por los ítems de selección fija.

Procedimientos estadísticos

Modelo de medición. La elección de cualquier modelo de medición está siempre limitada por las posibilidades que entregan los datos (por ejemplo, tamaño muestral, formato de los ítems y dimensionalidad). En el presente estudio se empleó un modelo de respuesta al ítem de la familia

Rasch para calibrar ítems y medir personas (Rasch, 1960; Wright, 1968; Wright & Masters, 1982). El Modelo de Crédito Parcial (Partial Credit Model, PCM) es una versión politómica del modelo Rasch (Fischer & Molenaar, 1995). Modela la probabilidad de pasar del nivel *j* al *j* + 1 dado que el examinado ya ha completado el paso del nivel *j* – 1 al j. Este instrumento posee cuatro niveles y cuatro parámetros de paso que se estimarán para cada ítem. Como se expresa formalmente en la Ecuación 1, para el PCM unidimensional, la probabilidad de que un examinado *n* complete el paso *j* para el ítem *i* es:

(1)

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x}(\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k}(\beta_n - \delta_{ij})},$$

donde β_n es el parámetro de habilidad de la persona *n* y δ_{ij} es el parámetro de paso del *j*-ésimo paso para el ítem *i* (Wright & Masters, 1982). Este modelo expresa la probabilidad de éxito de acuerdo a la diferencia entre la ubicación de la persona y la ubicación del ítem-paso. Los parámetros se estimaron con el software ConQuest (Wu, Adams, & Wilson, 1998). [8]

Los informes estadísticos generados por ConQuest se usan para describir estimaciones de los parámetros de persona e ítem y permiten investigar las propiedades de la escala CAL, incluyendo Teoría de respuesta al ítem y análisis tradicionales de ítems. También aplicamos análisis TRI estándar al ajuste de los ítems y las personas para evaluar el ajuste de los modelos. En la sección siguiente se presentan los resultados de los estudios de validez y fiabilidad aplicados a la parte de respuesta construida del instrumento CAL utilizando estos procedimientos psicométricos.

Resultados

Los resultados presentados en esta sección se apoyan en evidencia a favor y en contra de las inferencias sobre competencia personal (en este estudio, de profesores en formación) basadas en estimaciones obtenidas con la escala CAL. Se encontró evidencia que confirma nuestra hipótesis sobre la estructura unidimensional de la competencia de los profesores en formación con respecto a su comprensión del marco NRC y los tres dominios representados por el Triángulo de la evaluación (Figura 1). En primer lugar, investigamos la existencia de *niveles* distintos de comprensión sobre los temas incluidos en una versión modificada del Triángulo de la evaluación. En segundo lugar, examinamos cómo se puede distinguir significativa y sistemáticamente entre estos niveles de comprensión de los profesores empleando una escala psicométricamente sólida o un mapa Wright. En tercer lugar, presentamos evidencia con respecto a la fiabilidad y validez de las puntuaciones de la escala CAL y luego nos referimos a los diagnósticos y formativos apropiados del instrumento. Las implicaciones de estos hallazgos de acuerdo a las tres preguntas de investigación se tratan con mayor detalle en la sección de Discusión.

Los Estándares de evaluación (AERA, APA, NCME, 1999, 2014) guían la evidencia de fiabilidad y validez que se presenta en esta sección y además se usan para justificar las potenciales interpretaciones y usos del instrumento CAL. Se presentan cuatro evidencias de validez (*contenido*, *procesos de respuesta*, *estructura interna* y *relaciones con variables externas*) para apoyar la significación de las puntuaciones arrojadas por el instrumento CAL.

Primero, nuestra argumentación a favor de la *validez de contenido* se apoya en (a) el desarrollo del mapa de constructo que representa la intención de medir, (b) los ítems diseñados para estimular respuestas y (c) el espacio de resultados diseñado para valorar las respuestas de acuerdo al mapa de constructo (Wilson, 2005). El desarrollo de los mapas de constructo y los ítems del CAL se presenta como un ejemplo de este procedimiento de validación de contenido, el cual se llevó a cabo en un

período de tres años. Se halló evidencia adicional para apoyar "la relación entre el contenido del test y el/los constructo[s] que busca medir" (AERA, APA, NCME, 1999, p. 11) gracias a la revisión de la literatura y múltiples paneles para estudiar los ítems junto con académicos especialistas en metodología, profesores colaboradores y supervisores universitarios de la Fase II de la práctica profesional.[9]

Segundo, presentamos evidencia sobre validez basada en *procesos de respuesta* para comprobar "el ajuste entre el constructo y el carácter detallado del desempeño o la respuesta de los examinados" (AERA, APA, NCME, 1999, p. 12). Casi todos los 72 participantes completaron la entrevista de término del instrumento CAL. Los hallazgos globales obtenidos con las entrevistas de término fueron positivos, particularmente en el caso de los ítems de respuesta construida: "Se podrían aclarar los ítems de opción múltiple, por ejemplo el 2.8 y el 3.24". Otros participantes escribieron: "En general, la prueba fue muy buena", "se refirió al material del curso" y "los ítems [de respuesta construida] estaban escritos de manera clara". Sobre la base de los resultados de las entrevistas de término, concluimos que a la mayoría de los examinados no los confundió ni los distrajo el "ruido" externo (por ejemplo, carga de lectura, complejidad del lenguaje) que podría haber tenido un impacto negativo en su capacidad de responder los ítems de manera relevante de acuerdo al constructo (Messick, 1989).[10]

Tercero, se presenta evidencia sobre la validez de la interpretación de las puntuaciones de la escala CAL considerando la estructura y el funcionamiento de los 13 ítems de respuesta construida. Cuando se aplica un modelo de respuesta al ítem tipo Rasch para examinar la evidencia sobre la validez de la *estructura interna* de una escala, es importante presentar los resultados de los estadísticos de ajuste de medias cuadráticas ponderadas y *t*. Estos estadísticos de ajuste del modelo son una guía necesaria pero insuficiente para evaluar la evidencia a favor del uso de la escala en distintas situaciones, en este caso, en una evaluación diagnóstica. Nuestro análisis psicométrico de los estadísticos de ajuste de los ítems apoyan el hallazgo global de que los datos del instrumento CAL se ajustan bien al modelo de crédito parcial, lo que confirma la validez de su estructura interna.[11]

De acuerdo a los "Estándares de evaluación" (AERA, APA, NCME, 1999), la evidencia sobre la validez de la *estructura interna* se refiere al "grado en que las relaciones entre ítems y componentes se adecuan al constructo en que se basan las interpretaciones propuestas de un cierto instrumento" (p. 13). Sobre la base de nuestra manera de examinar la estructura de un constructo, un modelo de respuesta al ítem con un buen nivel de ajuste y capaz de describir diferencias cualitativas debiera establecer la distancia entre los examinados y las respuestas dentro de una escala. Nos interesa la probabilidad de dar una respuesta específica a un ítem/tarea en la escala CAL.[12]

Se utilizó un mapa Wright para examinar el ordenamiento empírico de personas e ítems y luego comparar estos valores con nuestras expectativas teóricas basadas en los mapas de constructo CAL. La Figura 6 muestra la distribución de las ubicaciones de los participantes y los ítems dentro de la escala CAL, considerando los resultados tanto pre-test como post-test.

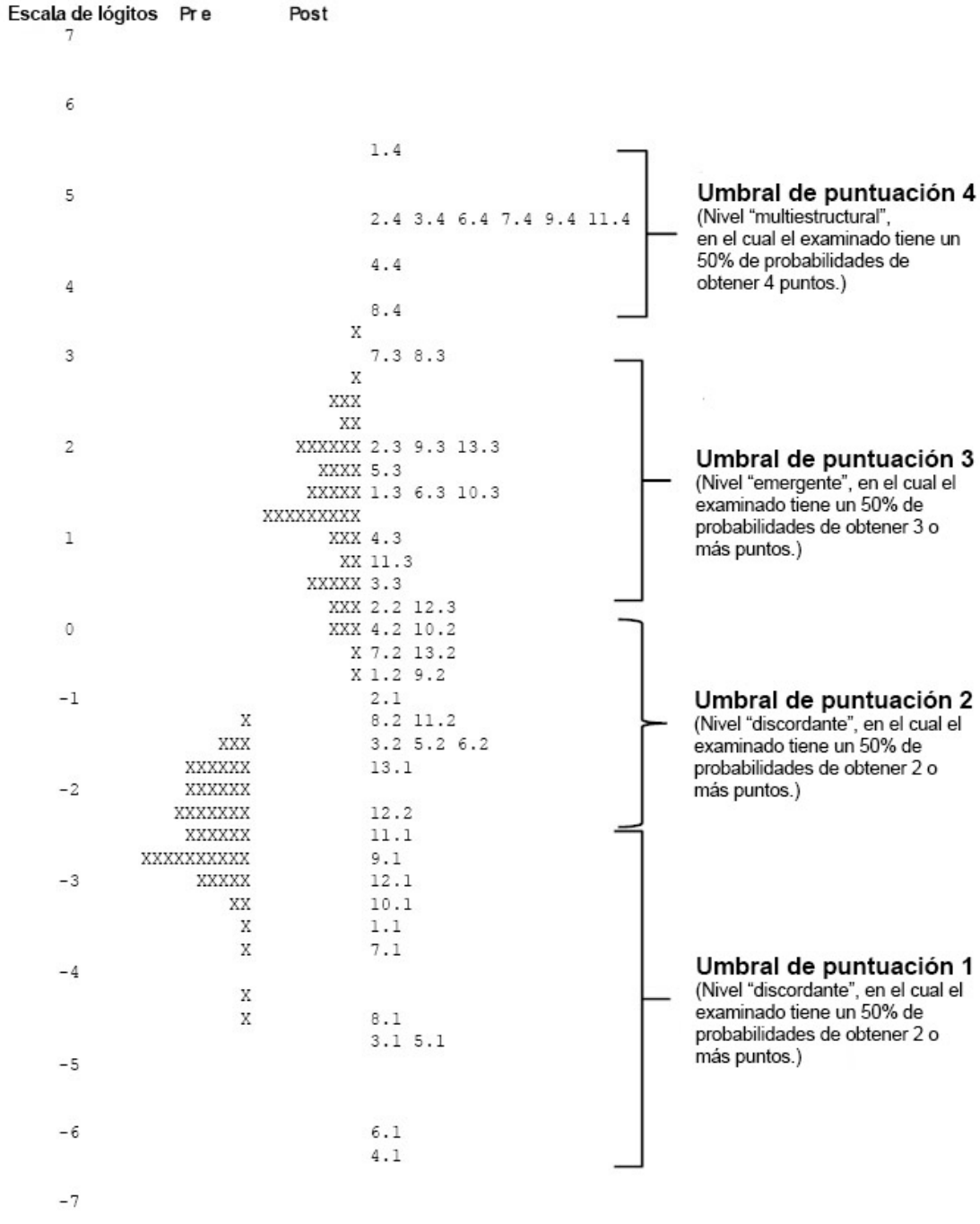


Figura 6. Mapa Wright pre- y post-test de competencias personales y umbrales de ítems para la escala CAL ('X' representa 1,7 casos).

Al comparar nuestra teoría sobre los constructos del CAL con el análisis de los datos empíricos del mapa Wright del mismo, hallamos evidencia que muestra un agrupamiento de los umbrales de los ítems, el cual parece indicar que [las respuestas a] los ítems dentro de un mismo nivel presentan dificultades similares en la mayoría de los casos. Como se muestra en la Figura 6, los umbrales de los ítems que representan los distintos niveles en la guía de puntuación del CAL (ver Figura 5) ocupan diferentes "bandas" dentro de la escala. Estas bandas están separadas, a excepción de un cierto solapamiento entre el primer y el segundo nivel. Específicamente, el nivel *preestructural* de respuesta a los ítems está representado por el primer umbral (es decir, los umbrales representados por "n.1", donde n va de 1 a 13), y cubre la parte inferior de la escala (de -6,55 a -2,66 lógitos). Los niveles de respuesta *discordante* y *emergente* están representados por el segundo y el tercer umbral (es decir, los umbrales representados por "n.2" y "n.3", respectivamente), los cuales ocupan la parte media de la escala (de -2,67 a 3,25 lógitos). Estos niveles representan niveles transicionales y emergentes de competencia en la respuesta a los ítems del instrumento CAL. Finalmente, observamos que el nivel *multiestructural* de respuesta, representado por el cuarto umbral (es decir, "n.4"), ocupa la parte superior de la escala (de 3,6 a 5,5 lógitos). Así, se puede afirmar que a medida que aumenta su competencia en el CAL, los participantes tienden a responder de mejor manera a la mayoría de los ítems, entregando respuestas más sofisticadas.

La segmentación de los niveles no es completamente limpia. Algunos umbrales de ítems, como por ejemplo MCA 2.1 y EID 13.1, se ubican en una región de solapamiento entre los niveles discordante y preestructural de competencia. Se requiere una investigación más específica de estos ítems, y tal vez algún tipo de modificación, para determinar si el problema se debe a los ítems o a una falta de diferenciación entre estos dos niveles. Por supuesto, se debe actuar con cautela al interpretar estas ubicaciones, ya que todas ellas son estimaciones con un error estándar asociado.

En general, se observa que los niveles del constructo CAL representados en la Figura 5 se reflejan en lo medular en los resultados empíricos de la Figura 6, aunque se mantiene algún grado de incertidumbre con respecto al límite entre los niveles preestructural y discordante. Asimismo, la relación entre profesores e ítems que se observa en el mapa Wright del CAL indica que la gama de competencias de los participantes queda totalmente cubierta por los umbrales de los ítems. No se detectó efecto techo ni suelo, lo que sugiere que el instrumento CAL captura bien los niveles de competencia de los examinados.[13]

Aunque no se muestran en una figura separada, los resultados empíricos de cada uno de los tres constructos parecen apoyar las expectativas teóricas presentadas en este estudio. En particular, se encontró que los umbrales de los ítems de las subescalas MCA, EHE y EID coinciden relativamente bien con la teoría de constructo (es decir, hipótesis con respecto a múltiples competencias en un modelo unidimensional). Los argumentos a favor de la validez de contenido de cada subescala arrojaron más evidencia sobre su alineamiento. Aunque determinar la estructura precisa y el funcionamiento exacto de cada uno de los constructos del CAL requiere una investigación que excede los límites de este estudio, en esta sección se presentan algunas observaciones preliminares basadas en los resultados de un estudio correlacional. La Tabla 2 muestra las correlaciones entre las subescalas del instrumento CAL (para 13 ítems de respuesta construida) empleando estimaciones de competencia derivadas de la aplicación del modelo de crédito parcial en ConQuest:

Tabla 2
Correlaciones entre subescalas según estimaciones de probabilidad ponderada del instrumento CAL

	Subescala MCA	Subescala EHE	Subescala EID
Subescala MCA		,761(*)	,817(*)
		,930	*,961*
Subescala EHE	,761(*)		,817(*)
	,930		*,942*
Subescala EID	,817(*)	,817(*)	
	,961	*,942*	

Nota. (*) indica que la correlación es significativa al nivel 0,001 (test de 2 colas). Las correlaciones desatenuadas aparecen en cursiva.

Los resultados que se presentan en la Tabla 2 indican correlaciones moderadamente fuertes entre subconstructos, que van de ,761 a ,817. Así, parece haber sólo un poco de evidencia que apoye la multidimensionalidad de la escala CAL dada las relativamente fuertes correlaciones entre mapas Wright calibrados por separado. Estos coeficientes de correlación son estadísticamente significativos y distintos de cero al nivel 0,001. Sin embargo, se debe tener cautela al interpretar las correlaciones de las estimaciones de parámetros de la teoría de respuesta al ítem que no consideren los errores estándar entre escalas o subtests. En general, un coeficiente de correlación se atenuará o reducirá debido a errores de medición; así, las estimaciones presentadas en la Tabla 2 pueden verse como límites inferiores de la verdadera correlación.

Si bien hay espacio para mejoras en la calibración y evaluación del diseño de ítems para este dominio específico, podemos concluir que entrega evidencia de validez relativamente buena a favor de la definición de constructo que coincide con la teoría general propuesta por el mapa de constructo CAL. Los resultados de un análisis general de los ítems fortalecen aún más la validez del instrumento en cuanto a su estructura interna. Wilson (2005) apunta que los problemas de validez de constructo están insertos tanto en el diseño de los ítems como en el mapa de constructo. Un requisito es que los ítems estén alineados con el instrumento visto como un todo. Específicamente, examinamos este punto en los ítems de respuesta construida estudiando las ubicaciones medias de cada grupo de puntuaciones dentro de cada ítem, las cuales debieran tender a aumentar a medida que suben las puntuaciones. En general, nuestra investigación sobre estas distintas líneas de evidencia sobre estructura interna nos lleva a concluir que el diseño de los ítems representa adecuadamente al constructo estudiado.

La cuarta fuente de evidencia sobre validez se basa en las relaciones entre el instrumento CAL y *otras variables externas*. Aquí examinamos "el grado en que estas relaciones se alinean con el constructo subyacente a las interpretaciones propuestas [del instrumento]" (AERA, APA, NCME, 1999, p. 13). Examinamos la relación entre las puntuaciones en el instrumento CAL y la Evaluación de desempeño para profesores de California (específicamente las rúbricas A6, A7 y A8, las cuales se refieren a la comprensión y uso de la evaluación en el aula por parte de los profesores en formación). El instrumento

PACT es una evaluación de desempeño pedagógico para que los profesores en formación reciban su licencia de Nivel 1. Fue desarrollado como una evaluación alternativa e "inserta en el currículo" apoyada en un "sistema basado en evidencias" para evaluar la preparación para enseñar (Pecheone & Chung, 2007).

En general, los coeficientes de correlación de Pearson entre estas puntuaciones PACT y las estimaciones de competencia personal del instrumento CAL mostraron una correlación positiva y moderada (*r*=,77). De hecho, la correlación mejoró al reexaminar la relación con las estimaciones de competencia personal considerando sólo los ítems de respuesta construida del instrumento CAL (*r*=,85). Esto indica que las puntuaciones del instrumento CAL podrían evaluar un conjunto de competencias referidas a la comprensión y el uso de evaluación en el aula similar al de las puntuaciones del instrumento PACT en el dominio de evaluación.

Los resultados del análisis de fiabilidad muestran la presencia de errores estándar de medición razonablemente pequeños. Esto se determinó calculando el error estándar promedio para personas en comparación con la gama completa de estimaciones theta para personas. La proporción entre ambos es de 16,03. Esto significa que, en esta muestra, nuestra comprensión de la competencia de los profesores en formación en cuanto a su alfabetización evaluativa es dieciséis veces más precisa que sin el instrumento CAL. También se observó que los coeficientes de fiabilidad de ambas versiones de la escala CAL alcanzaron valores altos. Para el instrumento completo, que incluye ítems de opción fija y de respuesta construida, los indicadores de consistencia interna como el alfa de Cronbach (,93) fueron altos. La fiabilidad de la escala CAL modificada, que incluyó solamente los 13 ítems de respuesta construida, fue ligeramente superior (,96).

Discusión y limitaciones del estudio

Este artículo presenta un conjunto de hipótesis sobre diferencias en el conocimiento sobre evaluación en el aula de un grupo de profesores, las cuales fueron probadas empíricamente mediante el análisis de las respuestas de 72 profesores en formación a un test pre-post, denominado *Instrumento CAL*. Si bien nos centramos en las propiedades técnicas de una escala piloto de Alfabetización Evaluativa en el Aula (Classroom Assessment Literacy, CAL), el estudio tiene implicaciones más amplias con respecto a cómo se entiende el sentido de la alfabetización evaluativa.

Como parte de esta investigación empírica, realizamos tres preguntas de investigación interconectadas con respecto a la estructura y función del constructo de alfabetización evaluativa. Para estudiar la primera pregunta ("¿Existen distintos niveles de comprensión sobre los temas del Triángulo de la evaluación?"), nos apoyamos en la evidencia de validez de contenido del instrumento CAL. A diferencia de estudios anteriores, y para especificar mejor el significado del constructo de "alfabetización evaluativa", desarrollamos mapas que describen la progresión de aprendizaje en los dominios temáticos principales. Todos estos temas están alineados con el marco NRC (2001a), un modelo mental empleado por expertos, el cual hace hincapié en el uso y la comprensión del *Triángulo del aprendizaje* en la ciencia y el diseño de evaluaciones desde el nivel preescolar al secundario.[14]

La segunda pregunta de investigación se basó en la primera. Si el contenido del instrumento CAL parece evaluar lo que busca evaluar, entonces "¿como distinguir significativa y sistemáticamente entre estos niveles de comprensión de los profesores en formación dentro de una escala psicométricamente sólida?" Inicialmente se abordó este tema aplicando un modelo de crédito parcial de respuesta al ítem tipo Rasch a los datos generados por el instrumento de 55 ítems mixtos para examinar nuestras

expectativas teóricas sobre la estructura de la competencia en el CAL. Descubrimos que los 13 ítems de respuesta construida, por sí solos, generaban un mejor ajuste en todos los dominios temáticos y aumentaban la fiabilidad del instrumento CAL. En otras palabras, los ítems de respuesta construida y sus correspondientes mapas Wright ofrecían mejores estimaciones de los niveles de desempeño. Confirmando nuestra teoría, estos mapas calibrados empíricamente se alinearon bien con los mapas de constructo.

Los análisis posteriores (es decir, los centrados en la asociación entre las subescalas MCA, EHE y EID) confirmaron un alto nivel de correlación entre los tres dominios temáticos.[15] Asimismo, encontramos una fuerte asociación positiva (r=,85) entre las estimaciones de competencia personal y las puntuaciones en el PACT (específicamente las puntuaciones promedio en las rúbricas A6, A7 y A8) de los mismos participantes en este estudio. Al parecer, el instrumento CAL, de modo similar al PACT Teaching Event, está detectando la comprensión y el uso de la evaluación en el aula por parte de los profesores en formación que buscan obtener su certificación para enseñar en California. Estos análisis estadísticos y psicométricos de teoría de respuesta al ítem entregan evidencia que demuestra la presencia de diferencias significativas y sistemáticas entre niveles de competencia (que van desde pre-evaluación a una comprensión más integrada del marco modificado del NRC) basadas en sus ubicaciones dentro de la escala CAL.

Aunque confiamos en que existe evidencia suficiente sobre validez y fiabilidad para apoyar ciertos diagnósticos o usos formativos del instrumento CAL, como por ejemplo en contextos de programas de formación de profesores, se debe tener mayor cautela con respecto a usos más amplios y no planificados.[16] En primer lugar, esperamos ampliar los formatos de los ítems y las plataformas de administración del instrumento CAL que se hallan actualmente disponibles. En particular, visualizamos oportunidades para obtener más datos a partir de plataformas tecnológicas mejoradas que permitan un mayor número de ítems de "limitación intermedia" (intermediate constraint) dirigidos a niveles de competencia específicos. En segundo lugar, es necesario realizar estudios de fiabilidad interobservador para examinar los ítems de respuesta construida empleando un modelo Rasch multifactorial (Linacre, 1989), especialmente si el instrumento se usa en varios lugares distintos, con diferentes evaluadores y en múltiples ocasiones. En tercer lugar, existen limitaciones potenciales para estudiar la estructura y funciones de las escalas CAL debido a las especificaciones de nuestro modelo actual de medición y el tamaño de la muestra (Brandt & Duckor, 2013).

El objetivo del presente artículo fue ampliar la visión que se tiene sobre la estructura del conocimiento acabado en evaluación y sobre cómo éste puede diferenciarse de las formas de entender la evaluación en el aula que caracterizan a los profesores principiantes. El estudio de las diferencias en el pensamiento individual de los profesores en formación con respecto a los bloques que conforman el diseño y el uso de la evaluación en el aula es parte de un proyecto de investigación a más largo plazo. Dicho enfoque tiene el potencial de ayudarnos a entender el crecimiento cognitivo de los individuos llamados "intermedios" (que no son ni expertos ni principiantes) y por lo tanto nos inspira como investigadores en educación y formadores de profesores. Es necesario seguir investigando las posibles progresiones de aprendizaje en los dominios de instrucción, currículo y evaluación en el caso de estos profesores en formación "emergentes", especialmente considerando su relación con las exigencias de contenido de cada asignatura. Ya es hora de investigar empíricamente, tanto en los años de formación pedagógica como durante el trabajo profesional en el aula, las progresiones de aprendizaje de los profesores en las diferentes líneas y dominios de la práctica de la evaluación.

El artículo original fue recibido el 15 de noviembre de 2016
El artículo corregido fue recibido el 22 de octubre de 2017
El artículo fue aceptado el 27 de octubre de 2017

Referencias

Adams, R. J., & Khoo, S. T. (1996). *Quest*. Melbourne, Australia: ACER.

Alonzo, A. C., & Gearhart, M. (2006). Considering learning progressions from a classroom assessment perspective. *Measurement: Interdisciplinary Research and Perspectives*, *4* (1&2), 99-108.

American Educational Research Association, American Psychological Association,
American Federation of Teachers, National Council on Measurement in Education, &
National Education Association. (1990). *The Standards for Competence in the Educational Assessment of Students.* Retrieved November 12, 2012, from http://buros.org/standards-teacher-competence-educational-assessment-students.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing (2nd ed.). Washington, DC: Author.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: Author.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In G. Sykes and L. Darling-Hammond (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3-32). San Francisco, CA: Jossey Bass.

Berliner, D. C. (1988, February). The development of expertise in pedagogy. Charles W. Hunt Memorial Lecture presented at annual meeting of the American Association of Colleges for Teacher Education, New Orleans, LA.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5,* 7-74.

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability.* Chicago: University of Chicago Press.

Borko, H., & Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American Educational Research Journal*, 26, 473-498.

Brandt, S. & Duckor. (2013, June). Increasing unidimensional measurement precision using a multidimensional item response model approach. *Psychological Test and Assessment Modeling, 55*(2), 148-161.

Braun, H. I., & Mislevy, R. (2005). Intuitive test theory. *Phi Delta Kappan*, *86*(7), 489-497.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006) Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*(1), 33-64.

Brookhart, S. M. (2001). *The Standards and classroom assessment research.* Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Dallas, TX. (ERIC Document Reproduction Service No. ED451189)

Campbell, C., Murphy, J. A., & Holt, J. K. (2002, October). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers.* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.

Corcoran, T.B., Mosher, F.A., & Rogat, A.D. (2009). *Learning progressions in science: An evidence-based approach to reform*. (CPRE Report). Philadelphia, PA: Consortium for Policy Research in Education.

Council of Chief State School Officers. (2010, July). Interstate Teacher Assessment and Support Consortium (InTASC) *Model Core Teaching Standards: A Resource for State Dialogue (Draft for Public Comment).* Washington, DC: Author.

Darling-Hammond, L. (2006). Assessing teacher education: The usefulness of multiple measures for assessing program outcomes. *Journal of Teacher Education*, *57*(2), 120-138.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work.* New York: Teachers College Press.

Duckor, B. (2005, May). *Thinking about the act of measuring: The development of a theory of the construct*. Individual poster presented at the 2nd annual meeting of the Center for As- sessment and Evaluation of Student Learning Conference, Santa Rosa, California. Available from Center for Assessment and Evaluation of Student Learning at http://www.caesl.org/ conference2005/ brent_sm.pdf

Duckor, B. M. (2006). *Measuring measuring: An item response theory approach*. (Doctoral dissertation, University of California, Berkeley, 2006). 345 pp. Advisor: Wilson, Mark R. *UMI Dissertation Abstracts (ProQuest).*

Duckor, B., Draney, K. & Wilson, M. (2009). Measuring measuring: Toward a theory of proficiency with the Constructing Measures framework. *Journal of Applied Measurement*, *10*(3), 296-319.

Duckor, B. (2014, March). Formative assessment in seven good moves. *Educational Leadership,* 71(6), 28-32.

Duckor, B., & Perlstein, D. (2014). Assessing habits of mind: Teaching to the test at Central Park East Secondary School. *Teachers College Record*, 116(2), 1-33.

Duckor, B. (October, 2017). *Linking formative assessment moves with high leverage instructional practices: Rethinking translation, application and practice of classroom assessment with a learning progressions framework.* Invited speaker at Graduate School of Education, Peking University, Beijing, China.

Duckor, B., & Holmberg, C. (2017). *Mastering formative assessment moves: 7 high-leverage practices to advance student learning*. Alexandria, VA: ASCD.

Dweck, C. S. (2010). Mind-sets and equitable education. *Principal Leadership*, *10*(5), 26–29.

Feiman-Nemser, S. (2001a). From preparation to practice: Designing a continuum to strengthen and sustain practice. *Teachers College Record 103*(6), 1013-1055.

Feiman-Nemser, M. (2001b). Helping novices learn to teach: Lessons from an exemplary support teacher. *Journal of Teacher Education*, *52*, 17-30.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Gotch, C.M., & French, B. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice*, *33*(2) 14–18.

Guskey, T. R. (2006). Making high school grades meaningful. *Phi Delta Kappan*, *87*(9), 670-675.

Heritage, H.M., Kim, J., & Vendlinski, T. (2008). Measuring teachers' mathematical knowledge for teaching (CSE Technical Report). Los Angeles, CA: Center for the Study of Evaluation and National Center for Research on Evaluation, Standards, and Student Testing.

Herman, J. L., & Baker, E.L. (2005). Making benchmark testing work. *Educational Leadership, 63*(3), 48-54.

Linacre, J. M. (1989). *Many-faceted Rasch measurement.* Chicago: MESA Press.

Lortie, D. (1975). *Schoolteacher: a sociological study*. Chicago, IL: The University of Chicago Press.

Marzano, R. J., Pickering, D. J., & Pollock, J. E. (2001). *Classroom instruction that works: Research-based strategies for increasing student achievement*. Alexandria, VA: Association for Supervision and Curriculum Development.

Mertler, C. A. (2000). Teacher-centered fallacies of classroom assessment validity and reliability. *Mid-Western Educational Researcher, 13(4),* 29-35.

Mertler, C. A. (2003). *Classroom assessment: A practical guide for educators.* Los Angeles, CA: Pyrczak.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to Evidence Centered Design*. CRESST Technical Paper Series. Los Angeles, CA: CRESST.

Molenaar, P. C. M., Huizenga, H. M., & Nesselroade, J. R. (2003). The relationship between the structure of inter-individual and intra- individual variability: A theoretical and empirical vindication of devel- opmental systems theory. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development.* Dordrecht, the Netherlands: Kluwer.

National Board for Professional Teaching Standards. (2012) *The five core propositions*. Retrieved from the NBPTS website: http://www.nbpts.org/the_standards/the_five_core_propositio

National Research Council. (2001a). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J.W. Pellegrino, N. Chudowsky & R. Glaser, (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, D.C.: National Academy Press.

National Research Council. (2001b). *Tests and teaching quality*. Washington, D.C., National Academy Press.

National Research Council. (2010). *Preparing teachers: Building evidence for sound policy.* Committee on the Study of Teacher Preparation Programs in the United States, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Nitko, Anthony, J., and Brookhart, Susan M. (2006) *Educational assessment of students*, 5th Edition. Upper Saddle River, New Jersey: Merrill.

PACT. (2007) *A brief overview of the PACT assessment system*. Retrieved from the PACT website: http://www.pacttpa.org/_main/hub.php?pageName=Home

Pecheone, R.L., & Chung, R. R. (2007). *Technical report of the Performance Assessment for California Teachers (PACT): Summary of validity and reliability studies for the 2003-04 pilot year.* PACT Consortium. Retrieved on March 17, 2012 from http://www.pacttpa.org/_files/Publications_and_Presentations/PACT_Technical_Report_March07.pdf

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher, 6(1),* 21-27.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: a national survey. Educational Measurement: Issues and Practice, 12(4), 10–12.

Popham, W. J. (1997). What's Wrong—and What's Right—with Rubrics. *Educational Leadership, 55*(4), 72-75.

Popham, W. J. (2000). *Testing! Testing! What every parent should know about school tests*. Boston: Allyn and Bacon.

Popham, W. J. (2004). Why assessment illiteracy is professional suicide. *Educational Leadership, 62*(1), 82-83.

Popham, W. J. (2007a). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan, 89(2),* 146-150.

Popham, W. J. (2007b). *Classroom Assessment: What teachers need to know* (5th ed.). Boston: Allyn & Bacon.

Popham, W. J. (2008). What's valid? What's reliable? *Educational Leadership, 65*(5), 78-79.

Putnam, R.T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29*(1), 4-15.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* [Reprinted by University of Chicago Press, 1980].

Shavelson, R.J, Moss, P., Wilson, M., Duckor, B., Baron, W., & Wilmot, D. (May, 2010). *The promise of teacher learning progressions: Challenges and opportunities for articulating growth in the profession.* Individual paper presented at the Teacher Learning Progressions symposium for Division D-Measurement and Research Methodology, American Education Research Association, Denver, Colorado.

Shepard, L.A. (2000). Role of learning in an assessment culture. *Educational Researcher*, *29*(7), 4-14.

Stiggins, R.J. (2001). The unfulfilled promise of classroom assessment. *Educational Measurement and Practice*, *20*, 5-15.

Stiggins, R.J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan*, *83*, 758-765.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, *77*(3), 238-245.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. (Edited by M. Cole, J. Scribner, V. John-Steiner, & E. Souberman). Cambridge, MA: Harvard University.

Wiggins, G. (1998). *Educative assessment*. Designing assessments to inform and improve student performance. San Francisco, CA: Jossey Bass.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New York, NY: Psychology Press.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, 46, 716-730.

Wright, B. D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing* (pp. 85-101). Princeton, NJ: Educational Testing Service.

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), *Measurement and multivariate analysis* (Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12-14, 2000), pp 325-332. Tokyo: Springer-Verlag.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wu, M., & R. J. Adams. (2012). Properties of Rasch residual fit statistics. Journal of Applied Measurement, 14, 339-355.

Wu, M. L., Adams, R. J., & Wilson, M. (1998). ACER*ConQuest* [computer program]. Hawthorn, Australia: ACER.

Endnotes

1 La palabra "distintos" se refiere a que los niveles difieren empíricamente de acuerdo a datos de puntuación calibrados. Esto deja abierta la pregunta de si estos niveles representan una teoría de constructos más fuerte, como por ejemplo niveles de desarrollo distintos entre personas. En relación con aquello, en el presente estudio se presume que cada profesor probablemente vaya alcanzando niveles cada vez más sofisticados de diferencias intrapersonales en cuanto a la sofisticación de su entendimiento; sin embargo, los datos son transversales y no pueden usarse para apoyar teorías referidas a estas diferencias a través del tiempo. Por ejemplo, véase Molenaar, Huizenga, & Nesselroade (2003).

2 Stiggins (1995) sostiene que "Las personas que cuentan con alfabetización evaluativa conocen la diferencia entre una evaluación sólida y otra que no lo es. No las intimida el mundo técnico de la evaluación, a veces misterioso y siempre abrumador" (p. 240).

3 Durante los últimos 20 años se han realizado estudios acerca de uno o más de los siete Estándares pedagógicos de competencia sobre la evaluación educativa de los estudiantes [Standards for Teacher Competence in the Educational Assessment of Students] (AFT, NCME, & NEA, 1990). Estos estudios muchas veces incluyen instrumentos de encuesta referidos a una amplia gama de acciones, desde el desarrollo de procedimientos de calificación válidos hasta el reconocimiento de prácticas poco éticas o ilegales (Brookhart, 2001; Plake, 1993; Plake, Impara, & Fager, 1993; Campbell, Murphy, & Holt, 2002). Campbell et al. (2002) administraron el Inventario de alfabetización evaluativa [Assessment Literacy Inventory, ALI] a 220 estudiantes de pregrado que se encontraban tomando un curso sobre tests y medición. El curso involucraba, entre otras actividades, crear y criticar diferentes métodos de evaluación, discutir sobre consideraciones éticas relativas a la evaluación, interpretar y comunicar resultados de evaluaciones estandarizadas y de aula y debatir y examinar las propiedades psicométricas (es decir, validez y fiabilidad) de las evaluaciones. Mertler (2000, 2003) empleó el Inventario de alfabetización sobre evaluación en el aula [Classroom Assessment Literacy Inventory, CALI] para comparar el nivel de "alfabetización evaluativa" de profesores en formación y en servicio. Las puntuaciones estándar y totales de los dos grupos de profesores fueron comparadas mediante pruebas t para muestras independientes (a = ,05). Los resultados mostraron diferencias significativas entre los dos grupos en 5 de los 7 Estándares, así como en las puntuaciones totales. Los profesores en formación tuvieron un mejor desempeño en el Estándar 1 (Elección de una evaluación apropiada).

4 Se identificó además un potencial cuarto dominio, Comprensión de los principios de la evaluación en el aula (PEA), el cual se refiere a las conexiones existentes entre estos cuatro dominios temáticos, particularmente con respecto a las nociones de validez y fiabilidad como pruebas de control de calidad en evaluaciones a nivel de aula, distrito y estado. El mapa representa una competencia global, entendiéndose el Triángulo de evaluación como un método de investigación que sirve para evaluar los juicios de otros evaluadores sobre los avances de los estudiantes. El mapa también considera cómo los profesores comunican las generalidades de las bases científicas y el diseño de las evaluaciones a otros actores interesados, incluyendo a padres, colegas, funcionarios administrativos y los propios estudiantes. El dominio PEA no se explora en este estudio. Puede encontrarse una consideración más general de este dominio en Duckor & Holmberg (2017).

5 En el presente estudio, dejamos abierta la pregunta de si las variables MCA en los niveles más altos deben ser específicas para cada dominio. Las habilidades y los conocimientos necesarios para evaluar materias como historia o álgebra probablemente requieran "niveles" más diferenciados y específicos para cada disciplina, así como probablemente nuevos mapas de constructo (por ejemplo, véase Wilson, 2009).

6 Una respuesta discordante típica en el análisis de una tarea de evaluación en matemática, por ejemplo, combinará la alfabetización en lectura con la fluidez procedimental: el profesor quiere evaluar ambas metas de aprendizaje pero aún no consigue definir cómo representarlas por separado en toda su complejidad.

7 Dado que el autor principal fue el único observador en el presente estudio, es posible que existan efectos de confusión intra-observador producto de su juicio tolerante/estricto con respecto a la dificultad de los ítems/pasos y los estadísticos de ajuste. Es necesario realizar estudios de fiabilidad interobservador para usos de datos de puntuación que involucren decisiones consecuentes o programáticas. Existe suficiente evidencia de fiabilidad sobre los usos del Instrumento CAL definidos actualmente, como por ejemplo diagnósticos a nivel de aula y evaluaciones formativas.

8 Se empleó la Estimación de probabilidad ponderada (Weighted Likelihood Estimate, WLE) como método específico de estimación de personas en todos los análisis de este artículo. Este método entrega las mejores estimaciones individuales con el menor grado de sesgo (Wu, Adams, & Wilson, 1998).

9 Sobre la base de investigaciones previas sobre la estructura de la competencia en el campo de la medición educacional (Duckor, 2006; Duckor, Draney, & Wilson, 2009), se entrevistó a expertos en contenido, académicos especialistas en métodos universitarios e instructores del programa de formación de profesores. Este enfoque basado en la teoría fundamentada (Grounded Theory) arrojó una imagen de posibles progresiones de aprendizaje, la cual posteriormente se convirtió en un conjunto inicial de mapas de constructo CAL. Estos mapas fueron corregidos luego de varias fases de pruebas piloto, excluyéndose algunos ítems en favor de otros para capturar mejor los niveles de desempeño, especialmente en el rango medio. A continuación triangulamos las respuestas a los ítems derivados de los instrumentos CAL piloto con un examen de escritos de los estudiantes del curso EDSC 182, lo cual permitió realizar mejorar adicionales a los mapas de constructo CAL, particularmente al mapa MCA, que se centra en la comprensión de la cognición de los estudiantes y la representación de metas de aprendizaje.

10 En general, las confusiones de los participantes tuvieron que ver con los ítems de respuesta fija. Sus comentarios se centraron en los distractores que “no tenían una respuesta clara” o entregaban “opciones confusas”. Un participante escribió lo siguiente: “Yo haría que algunas de las respuestas de opción múltiple fueran más finitas. Siempre tengo la impresión de estar respondiendo mal cuando hay tantas respuestas plausibles... parece que [estas preguntas] nos tratan de engañar”. Los participantes entregaron retroalimentación detallada sobre el instrumento, incluyendo sugerencias para mejorarlo. La mayoría de estas sugerencias se referían a reducir el tiempo necesario para completar el instrumento. Un profesor en formación comentó lo siguiente sobre el principio de evaluación encarnado en el instrumento: “El test es demasiado largo y va más allá del punto en que aumenta la fiabilidad”.

11 Los resultados del análisis de ajuste de los ítems coinciden con el hallazgo global de que, a nivel de ítems, los datos del instrumento CAL se ajustan razonablemente bien al modelo de crédito parcial. Examinamos estadísticos de ajuste de medias cuadráticas ponderadas (Wright & Masters, 1982) para los parámetros de paso de los ítems, los que indican un buen ajuste global según el marco interpretativo (,75 < MNSQ < 1,33) desarrollado por Adams y Khoo (1996). Solamente un ítem, el EHE 2.2 (categoría de puntuación 3) parece mostrar un desajuste con respecto al modelo desde un punto de vista estadístico. El valor de medias cuadráticas ponderadas de este ítem (1,28) es superior a 1, lo que indica que la varianza observada es mayor de lo esperado. Considerando el valor de t (2,4),

esto podría deberse no sólo al azar. Debe tenerse en cuenta que, dado el reducido tamaño muestral del estudio en comparación con el alto número de parámetros estimados, algunos parámetros (y por consiguiente los estadísticos de ajuste) podrían estar inflados debido a la tasa de errores tipo I. Con respecto a este tema, puede consultarse la discusión publicada por Wu y Adams (2012).

12 Los mapas Wright son herramientas visuales útiles para representar estas relaciones ("distancias") entre la competencia de las personas y la dificultad de los ítems dentro de un único continuo ("escala de lógitos"). En el mapa Wright presentado en la Figura 6 se interpretan las distancias entre profesores "X" y umbrales "*i.k*", donde *i* indica el ítem y *k* indica el umbral (por ejemplo, ver Wilson & Draney, 2002). Primero, si un profesor específico está en la misma ubicación que el umbral de un ítem del test CAL, como por ejemplo el umbral 2, esto significa que hay un 50% de probabilidades de que este profesor obtenga una puntuación "2" o inferior en dicho ítem. Segundo, un profesor más competente tendría una probabilidad superior al 50% de obtener una puntuación sobre "2" en ese ítem específico. Tercero, un profesor menos competente tendría una probabilidad superior al 50% de obtener una puntuación bajo "2" en ese ítem específico.

13 Debe tenerse en cuenta que nuestro diseño de ítems actual aún no entrega suficientes oportunidades para que los examinados demuestren un nivel integrado de comprensión en el instrumento CAL, asunto que se comenta en la sección de discusión.

14 Mediante una revisión de la literatura y entrevistas informales a profesores en formación, profesores colaboradores y varios supervisores universitarios de Fase II durante un período de dos años, se determinó que estos constructos alineados con el NRC podrían servir como punto de partida para insertar un test en el currículo a la vez que se cumplen metas más amplias relativas a los programas de formación y a las certificaciones para asignaturas específicas. De este modo, las evidencias presentadas en este estudio a favor de la validez de contenido del instrumento se apoyan en la relación entre estos mapas de constructo, los ítems que buscan fomentar respuestas en los profesores y las estrategias de puntuación para categorizar resultados derivados del instrumento, todos ellos elementos que se insertan en un currículo particular.

15 Las correlaciones fuertes (desatenuadas) entre las tres variables parecen desaconsejar la idea de tratar a estas progresiones de aprendizaje como subdimensiones separadas para fines de la administración de tests y evaluaciones. Sin embargo, desde un punto de vista educativo y pedagógico, los formadores de profesores y sus estudiantes podrían sacar provecho de esta distinción analítica, particularmente en cursos de evaluación en el aula enfocados en diseños de desempeño en tareas como los tests de unidades y los controles. Por ejemplo, véase Nitko & Brookhart (2006).

16 Las investigaciones dedicadas a medir la "alfabetización evaluativa" de los profesores, siempre que sea posible, deben indicar la confiabilidad de los instrumentos (por ejemplo, encuestas, herramientas de observación y tests) empleados para hacer afirmaciones sobre este constructo en poblaciones de profesores en formación. Nuestra revisión preliminar mostró que la escala CAL y sus subescalas tienen propiedades de confiabilidad razonablemente buenas. Además, los cuatro tipos de evidencia sobre validez recolectados en el estudio apuntan a una conclusión única: las relaciones entre los ítems del test y los componentes del mismo se ajustan al constructo en el que se basan las interpretaciones propuestas sobre las puntuaciones del instrumento CAL. Para una revisión exhaustiva de distintos instrumentos de "alfabetización evaluativa" y la evidencia técnica que justifica su uso, véase Gotwals & French (2014).