PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2013, 50(1), 4-20

Pontificia Universidad Católica de Chile
Facultad de Educación

# Combinación de mediciones de la práctica y el desempeño docente: consideraciones técnicas y conceptuales para la evaluación docente

# Combining Multiple Measures of Teacher Practice and Performance: Technical and Conceptual Considerations for Teacher Evaluation

**José Felipe Martínez**

Universidad de California, Los Ángeles, EE. UU.

## Resumen

Los esfuerzos recientes en reformas educativas en Estados Unidos y otros países prestan especial atención al desarrollo de sistemas de evaluación docente como complemento de los sistemas tradicionales de rendición de cuentas a nivel escolar, en la búsqueda por mejorar la práctica docente en el aula y el aprendizaje de los alumnos. En este artículo se examinan los aspectos conceptuales y metodológicos clave que se enfrentan al intentar medir constructos tan complejos y multidimensionales como la calidad y la eficacia docente, en el contexto de políticas de evaluación docente de alto impacto. Se consideran los métodos más comunes de medición de la práctica y el desempeño, así como los modelos disponibles para usar los diferentes indicadores en conjunto para propósitos de evaluación formativa y sumativa. El foco es la validez de las inferencias sobre la eficacia o calidad docente que se pueden derivar de las múltiples fuentes de información disponibles y las consideraciones, implicaciones y consecuencias potenciales para las políticas educativas una vez que estos sistemas se implementan.

***Palabras clave:*** evaluación docente, validez, mediciones múltiples, observación en clases

**Correspondencia a:**
José Felipe Martínez
Universidad de California, Los Ángeles, EE. UU.
2019B Moore Hall, Box 951521, Los Angeles, CA 90095-1521.
Correo electrónico: jfmtz@ucla.edu
Artículo originalmente escrito en inglés.



ISSN: 0719-0409 DDI: 203.262, Santiago, Chile
doi:10.7764/PEL.50.1.2013.2

## Abstract

Reform efforts in education systems in the United States and other countries place increasing emphasis on the promise of teacher evaluation systems to help complement traditional school-level accountability in improving classroom instruction and student learning. This paper offers a review of the key conceptual and methodological issues faced in measuring a construct as complex and multidimensional as teacher quality or effectiveness in the context of high-stakes accountability policies. The common methods of assessing teacher practice and performance are reviewed, along with approaches for using these different indicators in combination for formative and summative evaluation purposes. The focus here is on the validity of the inferences about teacher effectiveness that can be derived from these information sources, and the implications and potential consequences for education policy and the practice of implementing these systems in the field.

***Keywords:*** teacher evaluation, validity, multiple measures, classroom observation

## Evaluación docente y validez: consideraciones conceptuales y metodológicas en los sistemas de múltiples indicadores

> Entendemos a cabalidad que las pruebas estandarizadas no logran cubrir todas las sutiles características de una enseñanza exitosa. Es por esto que buscamos diferentes mediciones para la evaluación docente. En un mundo ideal, estos datos deberían servir para guiar la enseñanza y un desarrollo profesional útil. (Secretario de Educación de Estados Unidos, Arne Duncan, 2011).

La evaluación docente es un área que cada vez despierta mayor interés y reviste mayor importancia para los sistemas educativos locales, estatales y nacionales. Esto refleja un creciente consenso acerca de la necesidad de evaluar el desempeño de los maestros y de volverlos más responsables, en términos prácticos, de los niveles de logro académico de sus estudiantes. Se han puesto a prueba distintos métodos con el fin de evaluar las bases y los aspectos del desempeño docente. Además de los aspectos técnicos específicos de cada uno de los métodos, las preguntas que se encuentran al centro de los debates políticos sobre evaluación docente cada vez más se refieren a las formas más apropriadas de combinar la evidencia disponible de las distintas mediciones para formar juicios acerca del desempeño de los maestros.

Lamentablemente, aunque los legisladores y el público han prestado mayor atención, esta no siempre ha sido acompañada por un aumento en la conciencia y la comprensión de los complejos problemas técnicos que influyen en la evaluación del trabajo de los docentes. Existen pocas fuentes que puedan servir de guía a investigadores y legisladores para lidiar con la variedad de problemas conceptuales, técnicos y de políticas que se presentan al tratar de abordar este tema en la práctica. Por lo tanto, se necesita un gran trabajo sistemático y debate para responder las preguntas sobre los indicadores que se deben tomar en cuenta y las fuentes y usos apropiados de esa información, el grado de exclusividad y de traslapo de la información entregada por estas fuentes e indicadores y, lo más importante, las formas correctas de utilizar estos indicadores en combinación para evaluar a los maestros y mejorar su desempeño.

En el presente trabajo se examinan algunos de los conceptos claves y problemas metodológicos a los que se enfrentan los investigadores y legisladores que intentan medir la calidad o eficacia docente en el contexto de las últimas tendencias de la evaluación docente a nivel internacional. Específicamente, el objetivo de este trabajo es, por un lado, revisar la bibliografía sobre la evaluación con múltiples mediciones en los campos de la medición, la evaluación de estudiantes y la evaluación de personal y, por otro lado, traer este conocimiento al nuevo contexto conformado por sistemas de evaluación docente modernos a gran escala. El presente texto se divide en dos secciones: en la primera, se presentan los métodos más comunes que se utilizan para evaluar la práctica y el desempeño docente y se presentan ejemplos de su implementación en sistemas de educación de todo el mundo. En la segunda parte, se revisan los distintos modelos de medición que se proponen en la bibliografía para combinar los indicadores y se considera su implementación en los sistemas de medición múltiple de la evaluación docente. En especial, el texto se enfoca en la validez de las inferencias sobre la eficacia docente basadas en estos indicadores, ya sea por

separado o en combinación, y las posibles implicaciones y consecuencias de utilizar estos modelos en políticas educativas y en la práctica.

## Contexto de las políticas: por qué medir el desempeño docente

La teoría educacional, un volumen creciente de evidencia empírica y el sentido común sugieren que los maestros eficaces pueden ejercer una influencia importante en los niveles de logro de sus estudiantes (Baker et al., 2010; Rowe, 2003). Respectivamente, las actividades de reformas de los sistemas educativos en Estados Unidos y en otros países han dado mayor énfasis a la idea de la revisión del desempeño y la rendición de cuentas de cada maestro, en comparación o en conjunto con el foco tradicional de rendición de cuentas a nivel escolar. Al igual que en las olas de reformas anteriores centradas en la rendición de cuentas, esta nueva ola tiene su origen en las percepciones acerca del desempeño (inadecuado) de los estudiantes en evaluaciones nacionales o internacionales (por ejemplo, NAEP, PISA y TIMSS). Como describe Feuer (2012), ocurre un fenómeno interesante en muchos estados y países, donde los datos de estas evaluaciones generalmente se interpretan de forma pesimista con el fin de sugerir que el desempeño de los estudiantes es bajo en relación con otros alumnos o sistemas similares analizados.

Sin embargo, a diferencia de las reformas enfocadas a nivel escolar, estas iniciativas se apoyan sólidamente en supuestos y evidencia empírica sobre la importancia de la enseñanza de calidad y, de forma más general, la eficacia de los maestros (o la falta de eficacia) como factores que explican y, por lo tanto, se asocian a la mejora de los resultados observados. Además, reflejan una crítica que se aplica a la mayoría de los sistemas de evaluación docente, los que, como muchos informes y autores han señalado, se han vuelto rituales monótonos sin consecuencias ni utilidad y con poco valor ya sea sumativo, formativo o informativo, tanto para los maestros como para el distrito (Glazerman et al., 2010). Finalmente, estos datos reflejan una serie de supuestos optimistas sobre la capacidad de distinguir de forma confiable entre maestros eficaces e ineficaces que tendrían los sistemas de evaluación y los procedimientos actualizados que se han propuesto. Se puede observar claramente que estos razonamientos subyacen a las gestiones de reformas políticas de alta importancia en la actualidad, las que se enfocan en la evaluación docente en estados y distritos de Estados Unidos (por ejemplo, Los Ángeles, Chicago, Denver y Tennessee). El mismo discurso y los mismos supuestos se están volviendo más y más comunes en las actividades de reformas que apuntan a la evaluación docente a nivel internacional, donde se incluyen los sistemas nacionales como los de Singapur, Chile y México, y nuevas propuestas en el Reino Unido y Australia.

Aunque los sistemas de evaluación docente tienen diferentes orígenes, motivaciones y objetivos de acuerdo con el contexto de políticas del país, como grupo por lo general buscan recolectar información sobre aspectos diversos de la práctica docente dentro y fuera del aula, así como respecto del desempeño de los estudiantes, y pretenden utilizar la información recolectada para realizar evaluaciones formativas o sumativas. Estas actividades implican identificar a los maestros *con problemas* para intervenir y ayudarlos a mejorar su enseñanza y sus prácticas en clases; identificar a los maestros que generalmente presentan un mal desempeño para aplicar medidas correctivas o sanciones, o bien proceder al despido; brindar incentivos a los *mejores maestros*; proveer de información para las políticas a nivel de distrito y prácticas a nivel escolar sobre la formación de maestros y el desarrollo profesional; e identificar las prácticas eficaces con el fin de encontrar modelos de enseñanza eficaces que se puedan extender a las aulas de todo el sistema.

Generalmente, los sistemas de evaluación docente se crean con el fin de respaldar una combinación de objetivos formativos y sumativos, lo que implica alternar entre el uso formativo de la información recolectada para guiar la formación docente y el desarrollo profesional, y el uso sumativo para tomar decisiones relacionadas con el desarrollo profesional, el salario y la retención de personal. Es importante destacar que las prioridades tanto políticas como sobre políticas para la educación pueden oponerse al considerar con detención las implicancias y consecuencias de las diferentes opciones de la evaluación docente y la combinación apropiada de los objetivos formativos y sumativos para lograr un sistema a largo plazo. Por ejemplo, el epígrafe, de autoría del secretario Duncan (2011), refleja la determinación de seguir adelante con la evaluación docente de alto impacto, aun si las condiciones dictan que los datos de la evaluación no pueden utilizarse como retroalimentación para guiar las mejoras en términos de enseñanza y desarrollo profesional. Esto convierte el que debería ser el componente formativo esencial del sistema en un lujo opcional, que puede o no estar disponible en el *mundo real*.

## Qué evaluar

Aunque aumenta el consenso entre los investigadores, docentes y legisladores con respecto a la importancia de contar con enfoques razonables para la evaluacion docente, aunar criterios sobre los aspectos específicos de la práctica profesional docente a incluir en la evaluación es un tema de mayor complejidad. La *enseñanza* y la *práctica docente* constituyen en sí mismas constructos complejos y multidimensionales. La enseñanza comprende una variedad de procesos e interacciones dentro y fuera del aula; algunos son de naturaleza más sustantiva (contenidos y habilidades), otros se relacionan con los aspectos prácticos del trabajo en en aula (rutinas diarias y manejo de clase) y otros se vinculan a los aspectos psicológicos de las interacciones entre el maestro y el estudiante (por ejemplo, motivación, respeto y retroalimentación). La práctica docente, definida de modo más amplio, incluye también distintos aspectos del trabajo de un maestro fuera del aula, como por ejemplo reuniones con apoderados, administradores y otros maestros del colegio; deberes como *ciudadanos de un colegio* y contribuciones que aportan a la comunidad en general. Por lo tanto, aunque la noción de evaluar la *eficacia docente* tiene un atractivo intuitivo básico, en la práctica implica seleccionar, definir y recolectar información y extraer inferencias sobre docenas de constructos complejos (Peterson, 1987). Aunque términos tales como *calidad de la enseñanza*, *práctica docente*, *eficacia docente* o *desempeño docente* son usualmente tratados como intercambiables, es más apropiado verlos como términos que reflejan aspectos únicos de un constructo más amplio y, por lo tanto, presentan importantes áreas de intersección.

Por ende, evaluar la calidad o eficacia docente requiere una definición de cada uno de los constructos y componentes que han de medirse. La noción de *competencia docente* (Reynolds, 1999) brinda una heurística global útil para entender la calidad de la enseñanza de una manera amplia, que identifica cuatro componentes principales: el *conocimiento* docente (por ejemplo, conocimiento pedagógico y de la asignatura que se enseña), *habilidades* (por ejemplo, conocimiento aplicado), *disposición* (por ejemplo, actitudes, percepciones y creencias) y *prácticas* (por ejemplo, instrucción, asesoría y gestión). Vale destacar que esta definición excluye indicadores como la formación docente, títulos y credenciales profesionales, la experiencia (en la docencia y en grado) o los aportes al logro de los estudiantes en aspectos cognitivos y no cognitivos. A pesar de que estos indicadores por lo general se consideran en la evaluación formal de maestros, en el modelo de Reynolds se consideran más *correlativos* que *componentes* de la competencia docente.

También es importante destacar que no existe necesariamente un consenso con respecto a la definición precisa de cada uno de los componentes individuales en el modelo. Por ejemplo, aunque la noción de práctica en aula como un componente importante de la competencia docente puede tener amplio apoyo, todavía pueden existir desacuerdos entre distintas definiciones de los componentes específicos de la *prácticas docentes de calidad*. Para ilustrar este punto, en la Figura 1 se muestra el modelo de práctica en aula de calidad que se utiliza en Singapur en comparación con las dimensiones del modelo más utilizado en Estados Unidos (Danielson, 2007). Aun cuando la Figura 1 muestra algunas áreas que convergen, lo importante es que destacan la subjetividad y la especificidad cultural inherente a toda definicion de *buena enseñanza*. Por lo tanto, la observación en aula en Singapur, un país con alto rendimiento en las evaluaciones internacionales y cuyo sistema educativo ha sido objeto de mucho estudio y elogios en los últimos años, destaca aspectos de la vida en el aula que son poco considerados en los sistemas estadounidenses (por ejemplo, *enriquecer a los alumnos de forma integral para ganar su respeto y afecto*). Los sistemas estadounidenses, por otro lado, usualmente se centran en aspectos más *técnicos* de la práctica docente (por ejemplo, la claridad de la enseñanza, la planificación de clases, etc.).

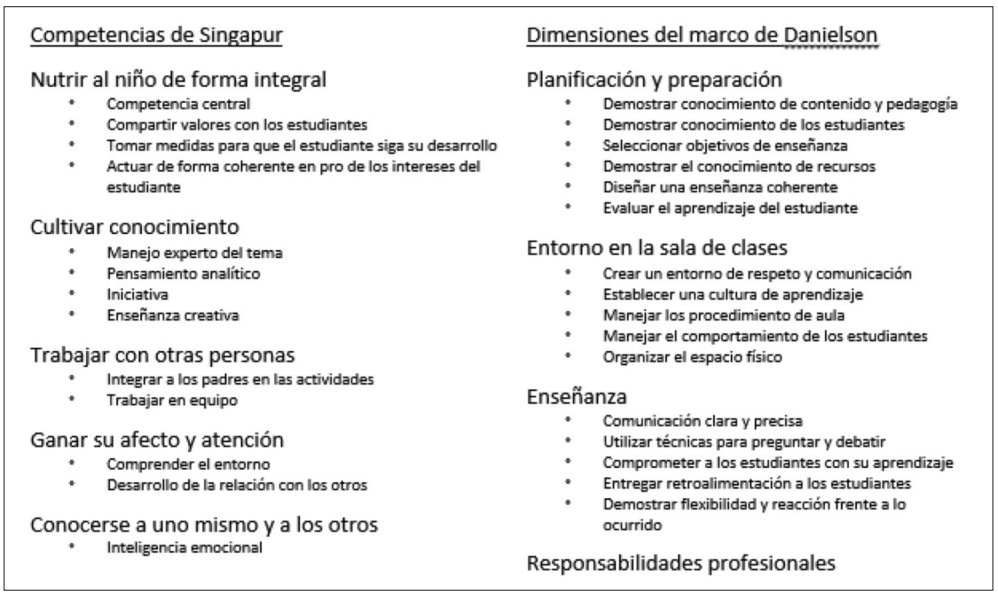


Figura 1. Constructos en Singapur y protocolos de observación de Danielson.

La discusión anterior sugiere, primero, que la definición más acertada de la calidad o competencia docente depende del uso que se le quiera dar y del contexto específico en que se sitúe este uso. Sin embargo, como regla general una respuesta sencilla a la pregunta sobre cuáles constructos evaluar cuando se pondera cómo medir de mejor manera el desempeño de los maestros podría ser «todos los anteriores» o, al menos, «todos los que sean posibles». Al excluir algunos de estos constructos de las definiciones de calidad o competencia docente, estas se acotan y, por ende, también lo hace la evaluación.

## Cómo evaluar

Una vez establecido que la evaluación docente debería incluir la mayoría de los aspectos o componentes críticos del desempeño docente desde la perspectiva conceptual y de políticas, se debe enfocar la atención en las preguntas metodológicas que se originan al medir cada uno de estos aspectos o componentes en la práctica. Se puede utilizar una gama de enfoques para recolectar información relacionada con cada constructo o aspecto de la calidad o del desempeño docente. En esta sección se presenta un resumen general de los métodos organizados dentro de las dimensiones del modelo de competencia docente de Reynolds (1989), en conjunto con la noción más reciente de eficacia docente en referencia al logro de los estudiantes. Cuando hay información disponible, se ofrecen ejemplos de sistemas en los cuales cada uno de estos métodos ha sido utilizado para la evaluación docente. Aunque estos ejemplos no son recomendaciones globales ni modelos ejemplares, pueden servir para ilustrar cómo se utilizan estos métodos para la evaluación docente en el mundo real.

**Pruebas del maestro**. Por lo general, el conocimiento y las habilidades de los maestros tanto de contenido como pedagógicos se miden por medio de pruebas estandarizadas (por ejemplo, la Praxis I) o evaluaciones de desempeño abiertas (por ejemplo, la evaluación Praxis II) o viñetas de práctica en el aula (Stecher et al., 2006). Últimamente ha aparecido un nuevo grupo de pruebas que buscan medir un tipo más particular de conocimiento pedagógico directamente ligado y basado en el conocimiento de contenido de asignaturas específicas. Entre los ejemplos de este tipo se pueden encontrar la prueba de conocimiento matemático para la enseñanza (Mathematical Knowledge for Teaching [MKT]) (véase Hill, Schilling, & Ball, 2004) y la prueba ATLAST para ciencias (Assessing Teacher Learning About Science Teaching ['evaluar el aprendizaje de maestros sobre la enseñanza de las ciencias']) (véase Smith & Taylor, 2010).

**Encuestas a docentes.** Las encuestas del maestro son métodos eficientes para recolectar información sobre distintos tópicos y constructos. Estas encuestas ofrecen indicadores que reflejan los constructos

clave de la práctica en aula, con niveles adecuados de confiabilidad en referencia a estándares sicométricos comunes. Sin embargo, las encuestas docentes tienen limitaciones importantes en su capacidad para producir información válida sobre los procesos en aula. Las encuestas están sujetas a errores de memoria y a incoherencias en las interpretaciones que los maestros hacen del contenido y el foco de los ítems. Por ejemplo, un maestro que indica *siempre* hacer hincapié en las habilidades cognitivas de orden superior en su enseñanza puede sobreestimar consciente o inconscientemente la frecuencia con que en realidad lo hace; de igual forma, los maestros pueden responder con alta sinceridad y precisión, pero entendiendo de manera diferente las palabras *siempre*, *hacer hincapié* u *orden superior*. Finalmente, las respuestas de los maestros pueden variar de acuerdo con su propia percepción de la necesidad o importancia de la práctica en cuestión (Mayer, 1999).

Las encuestas o entrevistas a docentes también se utilizan para recolectar información sobre disposiciones, creencias y percepciones (Mayer, 1999) y lo mismo ocurre con los constructos relacionados con el ambiente escolar y los aportes a la comunidad en general. Un ejemplo de lo anterior es la Encuesta de Condiciones Laborales (Working Conditions Survey, New Teacher Center, 2009), que se utiliza para medir las percepciones de los docentes, sus condiciones de trabajo y el clima escolar en distintos distritos de Estados Unidos.

**Encuestas a estudiantes.** Las encuestas a estudiantes se han vuelto más populares en los últimos años como una alternativa que puede abordar las limitaciones de las encuestas a docentes mencionadas anteriormente. En particular, las encuestas a estudiantes pueden utilizarse para crear indicadores al nivel del aula tan confiables como con las encuestas a maestros y con mayor poder predictivo para medir el logro escolar (Kane et al., 2012; Martínez, 2012). Dado que las encuestas a estudiantes están menos influenciadas por el contexto social, suele considerarse que son más válidas para propósitos de evaluación docente (Ferguson, 2010). Las encuestas a estudiantes brindan información adicional útil que usualmente no está al alcance de los maestros; en particular, se pueden utilizar las diferencias dentro de una misma clase en los informes de los estudiantes para supervisar la enseñanza diferenciada o individualizada (Martínez, 2012; Muthén et al., 1995). Al mismo tiempo, pedir información a los estudiantes presenta sus propios problemas en cuanto a la precisión y la coherencia de los informes, particularmente con niños más pequeños, mientras que con los niños mayores pueden presentarse sesgos. Las encuestas a los estudiantes también enfrentan desafíos metodológicos respecto del diseño y construcción de indicadores de los constructos correctos al nivel correcto (Schweigh, 2012). Por ejemplo, una pregunta que inquiere con qué frecuencia "mi maestro me pide que lea libros en clases" puede comportarse diferente en términos psicosométricos en comparación con una que inquiere con qué frecuencia "nuestro maestro nos pide leer libros". Finalmente, el uso de encuestas de alumno en la evaluación docente de alto impacto no ha sido probado empíricamente, por lo que puede traer complicaciones con respecto a su validez y parcialidad.

**Observaciones en clases.** A la larga, se ha utilizado una variedad de métodos para crear indicadores de prácticas en clases. Por lo general, estos se obtienen por medio de la observación del trabajo realizado por los docentes en sus propias aulas. La observación (ya sea en vivo o mediante grabaciones en video) tiene una validez considerable como un método para ver la enseñanza a medida que ocurre en las aulas. Asimismo, sirve como una evidencia directa para identificar áreas que deberían mejorarse, a fin de darles a conocer los programas de desarrollo profesional (Pianta & Hamre, 2009). Por otro lado, la observación en clases se enfrenta a importantes desafíos, que empiezan con identificar y definir los diversos aspectos de la *práctica* que se encuentran interconectados. En la Figura 1 se muestra un ejemplo de la complejidad y subjetividad relativa de los constructos implicados. Además, como un método de recolección de datos estandarizados a gran escala, la observación en clases se enfrenta a desafíos relacionados con la comprensión, cuantificación y monitoreo de la influencia del error humano a la hora de calificar las mediciones que se obtienen. Esta requiere técnicas de medición especializadas y una gran inversión en recursos para la capacitación, organización y monitoreo del trabajo que realizan los grupos de profesionales que trabajan como observadores estandarizados en un estado o distrito. Sin embargo, incluso con los recursos disponibles, todavía resulta difícil desarrollar y mantener mediciones confiables y a gran escala sobre las prácticas docentes. En un estudio reciente a gran escala en el cual se evaluaron algunas de las rúbricas de observación más conocidas, los resultados obtenidos recalcan sobriamente los desafíos a los que uno se enfrenta cuando se utilizan estas mediciones para respaldar inferencias y decisiones en relación a maestros individuales (Kane et al., 2012).

Las observaciones en clases siguen siendo el componente fundamental de los sistemas nuevos o antiguos

de la evaluación docente. Además, son el complemento central descriptivo y formativo de la evidencia sumativa obtenida de las mediciones basadas en los resúmenes de los logros escolares. Se puede ofrecer una variedad de ejemplos de observaciones en clases, incluidos los sistemas de evaluación docente rediseñados en Los Ángeles (Strunk, Weinstein, Makkonen, & Furedi, 2012) y Chicago (Sartain, Stoelinga, & Brown, 2011), además de otros ejemplos antiguos conocidos como los de Toledo y Cincinnati. El sistema nacional de desarrollo y de evaluación docente de Singapur es un ejemplo importante a nivel internacional (Sclafani & Lim, 2008). Los videos no han sido muy utilizados en la evaluación docente, pero se pueden encontrar ejemplos en los estudios recientes a nivel nacional en Estados Unidos (Ho & Kane, 2013) y, a nivel internacional, en el sistema de evaluación docente de Chile (Santelices & Taut, 2011).

**Portafolios.** Los portafolios de maestros son otro método que ha cobrado relevancia como una alternativa para recolectar información sobre las prácticas docentes. Los maestros utilizan portafolios para compilar, anotar y reflexionar sobre las modalidades de las prácticas docentes, como planes de clases, tareas y controles, a lo largo del tiempo. La literatura sugiere que los portafolios de maestros pueden utilizarse para recolectar información útil sobre las prácticas docentes a un nivel comparable con las observaciones en el aula (Martinez, Borko, & Stecher, 2011). Además, dado que requieren un esfuerzo y compromiso mental constante por parte de los maestros, los portafolios se consideran mecanismos útiles para el desarrollo profesional y para monitorear y mejorar las prácticas docentes (Shulman, 1998). No obstante, los portafolios requieren recursos considerables para su creación, recolección y revisión y pueden resultar una labor muy ardua para los maestros si no forman parte del ciclo de desarrollo profesional formal. Asimismo, los portafolios se ven limitados al capturar los aspectos interactivos o *verbales* de la enseñanza, como por ejemplo las sesiones de preguntas y respuestas que surgen de improvisto. Entre algunos de los prometedores ejemplos donde se han utilizado con éxito los portafolios de maestros, se encuentra el EdTPA, un sistema de evaluación basado en portafolios que se ha expandido rápidamente y que se enfoca en los estudiantes de pedagogía en práctica. Este sistema ha sido adoptado por 25 estados en Estados Unidos (Teacher Performance Assessment Consortium, 2012). Otro ejemplo son los portafolios integrales de maestros en el sistema de evaluación docente de Chile (Taut, Santelices, & Stecher, 2012).

**Modelos de valor agregado.** A pesar de que no forma parte del marco de *competencia* que se describe en párrafos anteriores, donde se considera un resultado o subproducto, el logro escolar es central para la idea de 'eficacia' que impulsa las reformas de políticas recientes sobre evaluación docente en Estados Unidos (no obstante, la evaluación docente también está conceptualmente vinculada con el logro escolar en otros países; el uso de los resultados de las pruebas rendidas por los estudiantes como una forma de evaluar a los maestros sigue siendo una estrategia particularmente utilizada en Estados Unidos hasta la fecha). Esto ha llevado al reconocimiento cada vez mayor de los modelos de *valor agregado* (Value Added Models [VAM]) para estimar la contribución de cada maestro al logro escolar, además de generar más debates en grupos de investigadores y legisladores. Se ha aludido a una variedad de problemas críticos en relación con las estimaciones provenientes de los modelos VAM, incluido su limitado alcance (Baker et al., 2010) y la falta de valor explicativo y de diagnóstico (Goe, Bell, & Little, 2008) para usos formativos, además de la inestabilidad (Schochet & Chiang, 2010) y la naturaleza no causal de los datos sumativos (Rubin, Stuart, & Zanutto, 2004). En conjunto, estas preocupaciones han motivado la noción de que los modelos VAM no pueden utilizarse por sí solos para evaluar a los maestros. Solo deben utilizarse junto a otras mediciones que incluyan un enfoque de evaluación más amplio (Braun, Chudowsky, & Koenig, 2010).

## Evaluación docente con múltiples mediciones

La multidimensionalidad intrínseca del desempeño docente como un constructo y las limitaciones de cada uno de los enfoques que se mencionan en este documento parecen indicar que ninguno de ellos es preferible en sí mismo o más útil en relación a los otros. Cada uno tiene sus ventajas y limitaciones, y cada uno resulta más adecuado para destacar diferentes aspectos importantes de la calidad o eficacia de los docentes. Por lo tanto, se deduce que ningún método por sí solo puede brindar información suficiente para respaldar una evaluación válida del desempeño de los docentes. Por el contrario, el consenso entre investigadores y legisladores es que la evaluación docente debe basarse en una gama de indicadores para que sea válida y útil (Baker et al., 2010; Braun, Chudowsky, & Koenig, 2010). Esta noción se ha utilizado ya por algún tiempo en las evaluaciones de estudiantes de alto impacto, como se refleja en los estándares para las mediciones psicológicas y educacionales:

> En los entornos educativos, no se debería tomar una decisión o dibujar una caracterización que influirá de manera sustancial [en los estudiantes] sobre la base de un resultado único. Se debería considerar otra información relevante si esta ayuda a mejorar la validez general de la decisión (AERA, APA, & NCME, 1999, p. 146).

Efectivamente, en Estados Unidos como en otros países, cada vez más distritos y estados se encuentran desarrollando o mejorando el enfoque que adoptan para la evaluación docente. Esto incluye buscar distintas fuentes de información para respaldar las inferencias evaluativas de alto impacto sobre los docentes cuando se deban tomar decisiones de contratación, ascenso y, en algunos casos, compensación. En Estados Unidos, la lista incluye a los tres distritos más grandes por numero de estudiantes (Nueva York, Los Ángeles y Chicago). En conjunto, estos tres distritos educan a más de dos millones y medio de estudiantes y emplean a más de ciento cincuenta mil maestros (Sartain, et al., 2011, para el caso de Chicago; Strunk et al., 2012, para Los Ángeles; y Marsh et al., 2011, para Nueva York). Otros ejemplos notables de esfuerzos recientes por mejorar la evaluación docente en Estados Unidos incluyen Washington D. C., Tennesee, Denver y Pittsburgh, entre otros.

Resulta interesante que, mientras la atención y el debate público respecto de estas políticas se ha centrado en el enfoque adoptado para estimar las contribuciones del maestro en el logro escolar (por ejemplo, los indicadores de valor agregado), en casi todos los casos el sistema asigna el mismo o mayor valor a otros indicadores del desempeño docente. Estos indicadores pueden incluir información obtenida de observaciones en el aula, informes de directores, encuestas a apoderados o estudiantes, instrumentos de enseñanza y registros oficiales, entre otros.

El creciente consenso en torno a este enfoque multidimensional se basa en una serie de suposiciones acerca de las consecuencias y beneficios de evaluar a los maestros utilizando indicadores múltiples. Entre otros beneficios, se espera que la evaluación docente con múltiples mediciones brinde una idea más completa del desempeño docente (Goe, Holdheide, & Miller, 2011), permita clasificar a los maestros en categorías más específicas y estables (De Pascale, 2012; Steele, Hamilton, & Stecher, 2010), minimice los incentivos para la preparación previa a las evaluaciones (Steele et al., 2010), brinde información para ayudar a los maestros a ajustar y mejorar sus metodologías pedagógicas y estrategias de enseñanza (Duncan, 2011) y genere más confianza en los resultados de las evaluaciones entre todas las partes interesadas, en particular los maestros y el público en general (Glazerman, Goldhaber et al., 2011). Estas y otras suposiciones han sido investigadas en el contexto de la evaluación del personal y las evaluaciones de estudiantes (Henderson-Montero, Julian, & Yen, 2003; Schafer, 2003). Sin embargo, no se comprende bien el alcance que cobran en conjunto cuando se aplican en la evaluación docente. Esto dependerá de varios factores, entre los que se incluye la naturaleza de los constructos involucrados, las inferencias a las que se quiere llegar, los usos de las mediciones y, aun más importante, los métodos específicos utilizados para *combinarlos* (Brookhart, 2009).

Este último punto adquiere particular relevancia si se considera el creciente interés en la evaluacion docente de alto impacto en Estados Unidos y en el mundo entero. Es interesante notar que el consenso en torno al uso de múltiples mediciones en este contexto es tan amplio como vago, pues las mediciones se pueden *combinar* de varias maneras para el propósito de evaluar maestros. La literatura especializada propone por lo menos cuatro enfoques distintos para combinar diversas mediciones que reflejen diferentes atributos de un constructo más amplio. Estos incluyen modelos de decisión no lineales conjuntivos y disyuntivos (o complementarios), y modelos compensatorios lineales para crear medidas o índices compuestos. En realidad, la mayoría de los distritos y estados han implementado una variedad de enfoques híbridos que combinan elementos de estos dos modelos base (Henderson-Montero, Julian, & Yen, 2003). No obstante, en el presente informe los modelos se analizan por separado para facilitar el análisis de las características particulares de cada uno. La pregunta crucial para los sistemas de evaluación docente es hasta qué grado el modelo escogido influye en las propiedades de los indicadores y, a fin de cuentas, las inferencias acerca de los docentes que se derivan de ellos (Mihaly, McCaffrey, Staiger, & Lockwood, 2013).

### Modelos conjuntivos

Este enfoque integra datos de diversas mediciones por medio de la determinación de una regla de decisión que requiere que los sujetos cumplan (*aprueben*) con un nivel mínimo de desempeño en cada

una de las mediciones pertinentes. Así, un modelo conjuntivo puede especificar que un maestro novato debe obtener puntajes de 3 o más en todas (o en cierto número especificado de) las escalas de medición de observación, que sea evaluado como «satisfactorio» en todas (u otro número especificado) las escalas de encuestas a estudiantes y que no quede en el 20% más bajo de la distribución de los puntajes de valor agregado para avanzar a una posición de docente titular. Se pueden especificar normas de decisión similares para otros propósitos (por ejemplo, para oportunidades de ascenso y mecanismos de incentivo) variando las mediciones incluidas y el desempeño estándar objetivo de cada una de las mediciones. La Figura 2 muestra de forma gráfica un modelo conjuntivo con múltiples mediciones, en donde se simplifica la evaluación a una serie de preguntas binarias para cada uno de los indicadores. Este modelo resulta más apropiado para minimizar *falsos positivos*, o en los casos en que es necesario un desempeño apropiado en áreas separadas o en componentes de un constructo más amplio. Dado que es más exigente que las otras alternativas, el modelo conjuntivo completo es poco común en su forma básica en los sistemas de evaluación docente a gran escala. Sin embargo, resulta interesante para comparar otros modelos y como parte de modelos complejos o híbridos.

### Modelos disyuntivos o complementarios

Estos son criterios de decisión específicos para la combinación de indicadores, donde lo que se busca es satisfacer los criterios de desempeño para un mínimo de *q* de *p* mediciones de indicadores en el sistema (Mehrens, 1989). En el ejemplo anterior, los maestros tendrían que cumplir con al menos dos de los tres criterios para aprobar una sección de evaluación dada. En la Figura 2 también se pueden apreciar los modelos de decisiones disyuntivas, ya que estos también se construyen al combinar una serie de decisiones específicas. El modelo disyuntivo sirve para los casos en los que se quiere reducir los falsos negativos, donde se recolectan mediciones repetidas del mismo constructo o en aquellos donde no todos los aspectos de un constructo más amplio tienen que conseguirse con la misma premura. Un caso especial del modelo disyuntivo que implica aprobar cualquiera de las medidas *p* se ha denominado *modelo complementario* (Brookhart, 2009) y se utiliza para maximizar la *oportunidad de aprobar* en situaciones de evaluaciones repetidas. Este tipo de modelo se utiliza en muchos sistemas modernos de evaluación docente, entre los que se incluyen Denver y New Haven en Estados Unidos y Singapur en el extranjero.

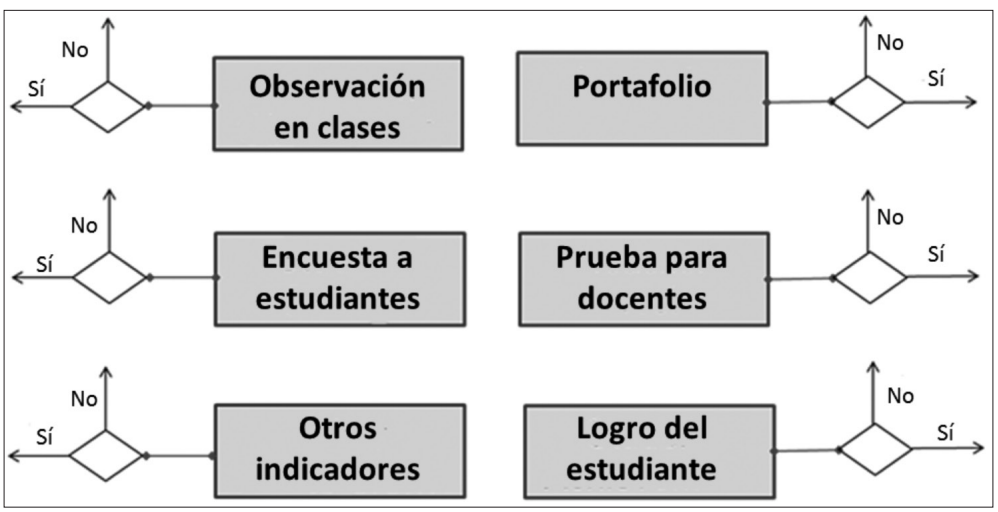


Figura 2. Modelos de decisión conjuntivos y disyuntivos.

**Modelos compensatorios**

En estos modelos, se permite que un buen desempeño en una o más mediciones compense un desempeño menos satisfactorio en otras mediciones. Esto se logra al crear indicadores compuestos o conjuntos que sintetizan la información disponible en todas las mediciones. En la mayoría de los casos, se asume adicionalmente que una única característica representa a los componentes subyacentes y se plantea un índice inclusivo a modo de promedio ponderado o simple (Brookhart, 2009). Alternativamente, los modelos compensatorios pueden ser considerados combinaciones lineales de mediciones que buscan maximizar ciertas propiedades de los indicadores obtenidos, o bien la relación que guardan con otras mediciones. Estos pueden incluir correlaciones empíricas entre diversos indicadores (modelos de análisis factoriales o canónicos) y relaciones entre los criterios medidos (Aamodt & Kimbrough, 1985) o no medidos (Darlington, 1970). En la Figura 3 se muestra una representación gráfica de un modelo compensatorio donde cada uno de los indicadores se concibe como un indicador o elemento dentro del análisis factorial. Este modelo permitiría investigar el patrón de interrelaciones que existe entre las mediciones y estimar cargas factoriales empíricas para crear un indicador compuesto que maximice la validez de las inferencias acerca de un constructo docente subyacente.

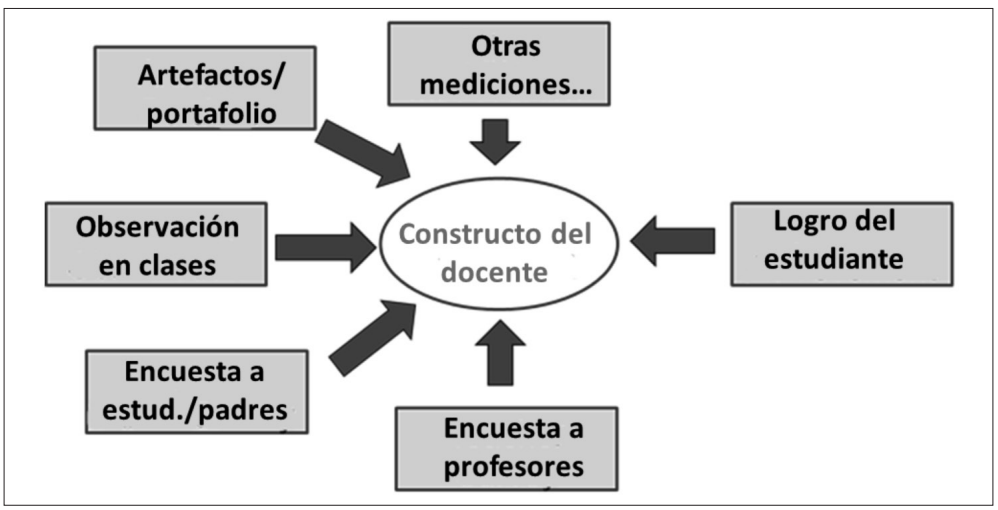


Figura 3. Modelo de análisis factorial para múltiples mediciones (compensatorio).

En la Figura 4 se detalla un modelo compensatorio diferente. Este es un modelo común en políticas e investigaciones que se enfocan a la rendición de cuentas para docentes, donde el logro estudiantil no se trata como un indicador, sino como el criterio de medición que los otros indicadores deben predecir. Por ejemplo, este es el modelo predeterminado en el reciente estudio MET sobre Mediciones de Enseñanza Eficaz (Measures of Effective Teaching), (Kane et al., 2012); en él se busca obtener parámetros de regresión (betas) para crear una combinación lineal óptima de indicadores que maximice el poder de predicción del criterio: logro escolar, ya sea estático o longitudinal como en los modelos de valor agregado. No obstante, al definir las evaluaciones estandarizadas como el criterio único, este modelo pareciera contradecir el objetivo principal de evaluar a los maestros sobre la base de múltiples mediciones que reflejen la naturaleza multidimensional de la profesión docente y los diferentes aspectos importantes del trabajo que los maestros realizan dentro y fuera de las escuelas y las aulas. Asimismo, dado que se asume que el criterio principal es conocido y está disponible, la inclusión de otros indicadores podría verse como innecesaria (pues sería redundante en relación con el criterio último con el que ya se cuenta) y podría incluso ser contraproducente (al diluir el criterio con indicadores indirectos de mas baja calidad).

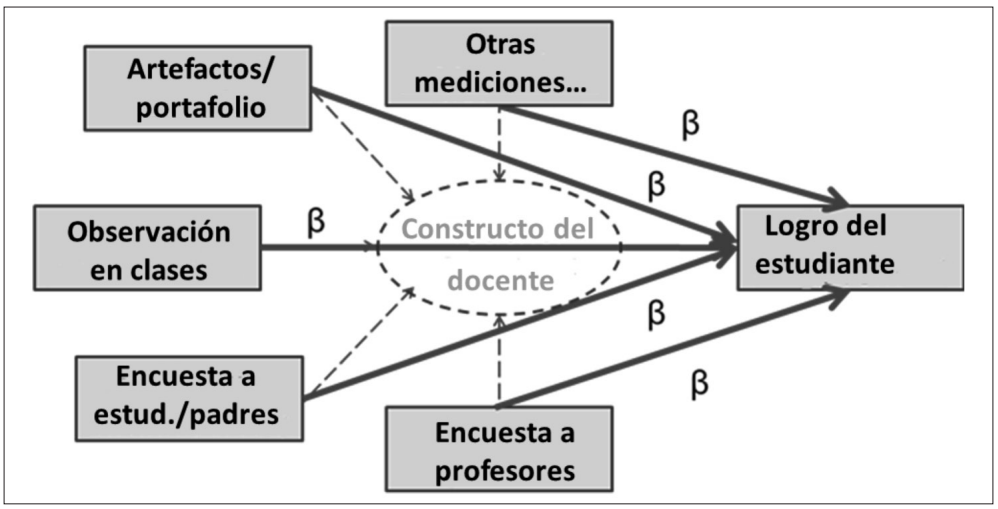


Figura 4. Ponderaciones de predicciones óptimas con los logros como criterio (compensatorio).

Finalmente, en la Figura 5 se presenta un modelo con un criterio no medido y ponderaciones teóricas relacionadas con los diferentes indicadores del desempeño docente (Darlington, 1970). Este modelo comparte características con los de las Figuras 3 y 4; al igual que el modelo de ponderación óptima, establece una regresión donde las ponderaciones beta especifican las relaciones entre cada indicador y criterio. Al igual que en el modelo de análisis factorial, el criterio es un constructo subyacente no medido que solo se puede inferir de manera indirecta a partir de los indicadores medidos incluido el logro escolar. La diferencia en el modelo descrito por Darlington (1970) es el enfoque que se adopta para determinar la ponderación más apropiada para cada medición. Esta es una pregunta cada vez más importante para los legisladores, que este modelo sugiere no se puede responder desde una perspectiva estrictamente técnica y científica; las ponderaciones apropiadas deben asignarse de manera *teórica*, es decir, a partir del consenso al que lleguen las partes involucradas claves respecto de las metas y prioridades de un sistema de evaluación y el valor que se asigna a cada una de sus partes. Este es el modelo que se está implementando por necesidad en muchos distritos, dado que la alternativa seria tomar decisiones cruciales sobre las oportunidades de ascenso o compensación para docentes con base en ponderaciones empíricas que pueden fluctuar entre distritos y dentro de los distritos a través del tiempo. Si se considera que el modelo elegido deberá presentarse eventualmente a los sindicatos de maestros y al público en general, se concluye que las ponderaciones teóricas están de hecho mejor adaptadas para utilizarse en evaluaciones docentes de alto impacto y gran escala (Darlingon, 1970).

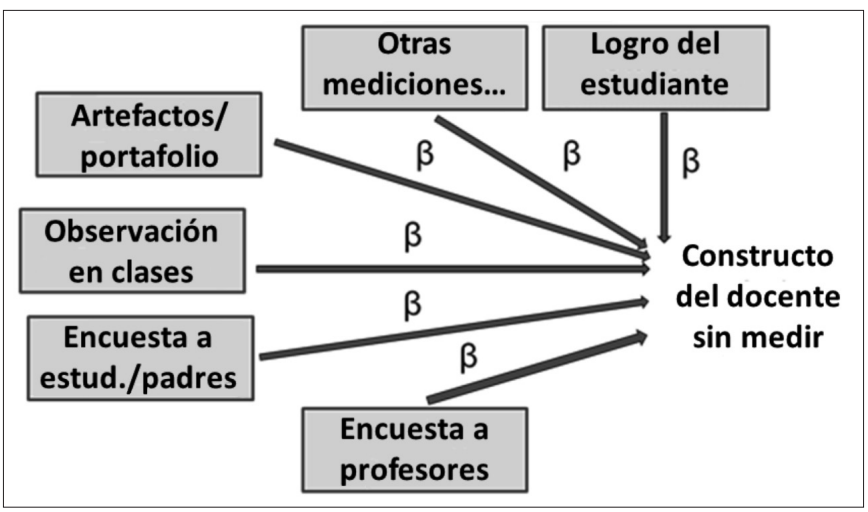


Figura 5. Criterio no medido con ponderaciones teóricas (compensatorio).

### Modelos híbridos

Los tres enfoques generales recién descritos pueden integrarse en modelos de evaluación híbridos, si se desea. Un modelo híbrido conjuntivo-disyuntivo podría especificar que todos los maestros deben cumplir con los criterios del logro escolar y al menos uno de los otros dos criterios adicionales (observación en clases y encuestas a estudiantes). Por su parte, un modelo conjuntivo-compensatorio podría requerir que los maestros cumplan con un conjunto de criterios, donde uno de ellos es un compuesto lineal de tres mediciones compuestas de encuestas a maestros y estudiantes.

Por ultimo, cabe destacar que un enfoque alternativo que por lo general se pasa por alto, se basa en la idea de que no es necesario *combinar* las diferentes mediciones, sino más bien deberían utilizarse *en conjunto* a fin de brindar información más sustancial y útil para respaldar una evaluación formativa (Schmidt & Kaplan, 1971). Tal como indica Mehrens (1989), la recolección, mantenimiento, informes y uso de las mediciones de manera separada parece ser la elección natural para las evaluaciones docentes, donde los propósitos clave se centran en procesos formativos y en el desarrollo profesional para mejorar la práctica docente. En un modelo formativo cada una de las mediciones guía y otorga información para los esfuerzos de mejora relacionados con los aspectos del desempeño y la práctica docente. Finalmente, si se requieren, los indicadores pueden aun combinarse para propósitos sumativos de acuerdo a cualquiera de los modelos de decision mencionados.

### Mediciones múltiples y confiabilidad

Los modelos conjuntivos y disyuntivos son intuitivamente interesantes y, en principio, pueden implementarse sin usar técnicas estadísticas avanzadas. Sin embargo, la confiabilidad de los indicadores e inferencias que se obtienen a partir de cada uno de estos enfoques puede variar de manera sustancial y debería examinarse de cerca para garantizar que el sistema ofrezca información que sea lo suficientemente precisa para el propósito que se busca. Específicamente, a pesar de la suposición común de que el uso de múltiples mediciones mejorará la confiabilidad de las inferencias, se puede mostrar que con el modelo conjuntivo la confiabilidad de una decisión es una de las menos confiables de las mediciones que lo componen (Chester, 2003). Supongamos, por ejemplo, que los *puntajes verdaderos* del maestro A cumplen con los criterios M1 y M2; el maestro debería *aprobar* ambas mediciones. Sin embargo, dado que las mediciones no son completamente confiables, el maestro A podría ser clasificado de manera inadecuada como *reprobado* en las mediciones. Si conociéramos los resultados reales del maestro A, podríamos estimar la probabilidad de clasificación errónea directamente desde el error estándar de medición de cada medición. Para el maestro A estas probabilidades hipotéticas se estiman en 0,25 y 0,15, respectivamente, para M1 y M2; es decir, la probabilidad de una clasificación correcta es de 0,75 y 0,85. A partir de esta información se puede deducir que la probabilidad de aprobar ambas pruebas en práctica (de acuerdo con lo observado, no en los resultados reales) con un modelo conjuntivo es de 0,75 * 0,85 = 0,64; mientras que la probabilidad de *aprobar* al menos una vez con el modelo disyuntivo es de 1 - [0,25 * 0,15] = 0,96. A pesar de que Douglas y Mislevy (2010) sugieren que las diferencias pueden no ser tan importantes con los modelos compensatorios, el ejemplo acentúa la importancia de comprender los errores de medición en los contextos de las evaluaciones con múltiples mediciones, particularmente aquellos que implican normas de decisiones complejas o híbridas (Cronbach, Linn, Brennan, & Haertel, 1997; Douglas & Mislevy, 2010).

### Mediciones múltiples y validez

Al igual que con la confiabilidad, en el caso de la validez solo nos encontramos en las etapas iniciales de los estudios sobre los diferentes enfoques para combinar la gran variedad de mediciones del desempeño docente disponibles. A medida que estos sistemas se utilicen más en línea y en terreno, mayor será la urgencia de contar con estudios empíricos sobre la validez a fin de investigar las suposiciones, implicaciones, requisitos y posibles consecuencias de utilizar cada uno de los modelos para formar juicios acerca del desempeño de los maestros (Brookhart & Loadman, 1992).

Como se mencionó anteriormente, la decisión sobre como combinar las diferentes mediciones del desempeño docente presenta problemas no solo técnicos sino también conceptuales e incluso de índole

politica. Los sistemas de evaluación docente primero deben explicitar lo que entienden por eficacia y definir el valor que otorgan a las mediciones basadas en nociones teóricas o empíricas, siempre a la luz del contexto local, los objetivos del programa y las prioridades. Cuando se consideran estos diferentes elementos y decisiones, resulta útil concebirlos como componentes de un argumento sobre la validez para inferencias específicas que se deduce de los indicadores para propósitos específicos (Kane, 2006). Como con cualquier medición por sí sola, el marco de validez que se aplica a las múltiples mediciones destaca la necesidad de confiar en suposiciones sobre la naturaleza del constructo teórico que se mide (es decir, calidad o eficacia docente) y las maneras en que se puede instaurar en la práctica. Asimismo, deja en claro que los diferentes usos llevarán a diferentes argumentos de validez y requieren diferentes fuentes de evidencia y apoyo. En particular, las inferencias y usos que acarrean consecuencias importantes requieren un mayor respaldo empírico y teórico.

Tal como lo menciona Kane (2006), la validez es unitaria y depende del propósito; la validación implica definir un argumento interpretativo para las inferencias, usos y consecuencias que se buscan y recolectar evidencia que respalde ese argumento. Esta evidencia puede incluir una variedad de fuentes de respaldo teórico para los constructos y marco conceptual esperados, además de evidencia empírica coherente y precisa (confiabilidad), patrones de intercorrelación y estructura interna para las mediciones y capacidad de predicción sobre los criterios de medición estudiados, entre otros. Asimismo, el marco también requiere una consideración explícita de las consecuencias y resultados que se espera obtener de las interpretaciones y usos propuestos de las mediciones. Es claro que la evidencia de consecuencias esperadas y no esperadas es de hecho la consideracion clave desde una perspectiva de políticas; de lo contrario, la validez se convierte en un tema más bien académico y menos relevante para los sistemas en operación.

## Discusión

### Consideraciones para evaluar a los docentes con el uso de múltiples mediciones

Una nueva generación de reformas que enfatizan la rendición de cuentas orientada a los docentes se está implantando en sistemas educativos de todo el mundo. El remplazo de rituales de evaluación superficiales por sistemas exhaustivos para evaluar y monitorear la enseñanza y desempeño docente parece ser prometedor para influenciar y mejorar la práctica docente y resulta interesante para los legisladores y el público en general. La nueva camada de sistemas de mediciones múltiples tiene el potencial de convertirse en el motor que impulse el cambio hacia una cultura de reflexión, desarrollo y rendición de cuentas entre los maestros y de respaldar los sistemas de evaluación formativa cuyos resultados se integran directamente a los canales de desarrollo profesional, lo que lleva a mejoras significativas en la enseñanza y, a la larga, al aprendizaje de los estudiantes. No obstante, estos sistemas se enfrentan a desafíos conceptuales y metodológicos considerables relacionados con su diseño e implementación y evolucionan insertados dentro de entornos de políticas complejos.

La eficacia docente es un constructo multidimensional complejo y la evaluación docente depende en gran medida de los propósitos y el contexto; la intención de este articulo no es ofrecer recomendaciones generales a los encargados de la formulación y creación de políticas sino delinear las consideraciones técnicas y conceptuales claves a tomar en cuenta en el diseño de tales sistemas. Idealmente, estos deberían construirse sobre la base de un modelo de enseñanza eficaz que defina los constructos de conocimiento, habilidades y conductas que se incluyen en la noción más amplia de calidad o eficacia docente. Este modelo es crucial porque rescata tanto los componentes sumativos de la evaluación como los componentes de desarrollo profesional formativos. De esta manera, se pueden deducir de manera explícita y modelar las expectativas y mejores prácticas para aspectos cruciales del trabajo de los maestros.

Además de considerar la definición de cada constructo y la calidad de los indicadores disponibles, los sistemas de educación deberían considerar que, como se ha comentado en este documento, la combinación de múltiples indicadores falibles no resulta automáticamente en *mejores* inferencias pero menos falibles, sino más bien ofrece inferencias *más complejas*. Específicamente, los legisladores deberían considerar detenidamente las suposiciones y consecuencias (esperadas e inesperadas) de los diversos enfoques para combinar mediciones o para utilizarlas en conjunto. Asimismo, la evaluación docente incluye inferencias sobre maestros en específico y, por lo tanto, no se puede definir en los resultados en términos de productividad y otros parámetros econométricos. Dado que el impacto para los maestros es alto, es importante que los desarrolladores se centren en la validez de los usos e inferencias esperados y tomen las medidas necesarias para desarrollar un sistema que respalde apropiadamente estas inferencias. Por ende, los sistemas de múltiples mediciones deben examinar de manera explícita y atenta cómo los diferentes modelos para combinar mediciones influyen en las inferencias que se deducen sobre los docentes (Mihaly et al., 2013).

Finalmente, para evitar que se conviertan en una "varita mágica" que se encienda y esfume con rapidez, o se diluya hasta perder su sentido, los sistemas de múltiples mediciones de evaluación docente deberían contar con un respaldo teórico y empírico suficiente a lo largo de su desarrollo e implementación. Como investigadores y metodólogos, es nuestro trabajo recordar a los legisladores que toma tiempo lograr buenas mediciones, que toma más tiempo evaluar e implementar los sistemas sólidos basados en estas mediciones y que las consecuencias de los usos específicos de estos sistemas en gran medida no se conocen y, por lo tanto, su evaluación tomará más tiempo. El mayor riesgo no se encuentra en la posibilidad de decisiones injustas que impliquen a maestros específicos, aunque la posibilidad de que ocurran debería ser preocupante para los legisladores; más bien, el riesgo está en perder una oportunidad crucial de promulgar políticas correctas con el potencial de influenciar de manera positiva la práctica educacional y sus resultados.

El artículo original fue recibido el 11 de diciembre de 2012
El artículo revisado fue recibido el 12 de febrero de 2013
El artículo fue aceptado el 13 de febrero de 2013

**Referencias**


Aamodt, M. G., & Kimbrough, W. W. (1985). Comparison of four methods for weighting multiple predictors. *Educational and Psychological Measurement, 45*, 477-482.

AERA, APA, & NCME (1999). *Standards for educational and psychological testing.* Washington, D. C.: American Educational Research Association.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers*. Washington, D. C.: Economic Policy Institute.

Braun, H., Chudowsky, N., & Koenig, J. (Eds.) (2010). *Getting value out of value-added*. Washington, DC: National Academies Press.

Brookhart, S. (2009). The many meanings of multiple measures. *Education Leadership, 67*(3), 6-12.

Brookhart, S. M., & Loadman, W. E. (1992). School-university collaboration: across cultures. *Teaching Education*, *4*(2), 53-68.

Chester, M. D. (2003). Multiple measures and high-stakes decisions: a framework for combining measures. *Educational Measurement: Issues and Practice*, *22*, 32-41. doi: 10.1111/j.1745-3992.2003.tb00126.x

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*, 373-399. doi: 10.1177/0013164497057003001

Danielson, C. (2007). *Enhancing professional practice: a framework for teaching.* Alexandria, VA: Association for Supervision and Curriculum Development.

Darlington, R. B. (1970). Some techniques for maximizing a test's validity when the criterion variable is unobserved. *Journal of Educational Measurement*, *7*, 1–14. doi: 10.1111/j.1745-3984.1970.tb00688.x

De Pascale, C. (2012). Managing Multiple Measures. *Principal*, *91*(5), 6-10.

Douglas, K. M., & Mislevy, R. J. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics, 35,* 1-27. doi: 10.3102/1076998609346969

Duncan, A. (2011). *Duncan tells teachers: change is hard.* Recuperado el 9 de diciembre de 2012 de http://www.ed.gov/blog/2012/08/duncan-tells-teachers-change-is-hard/

Ferguson, R. (2010). *Student perceptions of teaching effectiveness.* Recuperado de http://www.gse.harvard.edu/ncte/news/Using_Student_Perceptions_Ferguson.pdf

Feuer, M. (2012). *No country left behind: Rhetoric and reality of international large-scale assessment*. Princeton, NJ: Educational Testing Service.

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S., Staiger, D. O., & Whitehurst, G. J. (2011). *Passing muster: evaluating evaluation systems.* Washington, DC: Brown Center on Education Policy at Brookings.

Glazerman, S., Loeb, S., Goldhaber, D., Steiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: the important role of value-added*. Nueva York: Brookings. Recuperado de www.brookings.edu/research/reports/2010/11/17-evaluating-teachers

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: a research synthesis*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goe, L., Holheide, L., & Miller, T. (2011). *A practical guide to designing comprehensive teacher evaluation systems*. Washington, D. C.: National Comprehensive Center for Teacher Quality.

Henderson-Montero, D., Julian, M., & Yen, W. (2003). Multiple perspectives on multiple measures. *Educational Measurement: Issues and Practice*, *22*(2), 6. doi: 10.1111/j.1745-3992.2003.tb00121.x

Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal, 105*, 11-30.

Ho, A., & Kane, T. (2013). *The reliability of classroom observations by school personnel.* Seattle, WA: Bill & Melinda Gates Foundation. Recuperado el 2 de enero de 2013 de http://www.metproject.org/downloads/MET_Reliability%20of%20Classroom%20Observations_Research%20Paper.pdf

Kane, M. T. (2006). Validation. En R. L. Brennan (Ed.), *Educational measurement, Fourth Ed.* (pp. 17-64). Westport, CT: American Council on Education and Praeger Publishers.

Kane, T. J., Staiger, D. O., McCaffrey, D., Cantrell, S., Archer, J., Buhayar, S., Kerr, K., Kawakita, T., & Parker, D. (2012). *Gathering feedback for teaching: combining high-quality observations with student surveys and achievement gains*. Seattle: Bill & Melinda Gates foundation.

Marsh, J. A., Springer, M. G., McCaffrey, D. F., Yuan, K., Epstein, S., Koppich, J., Kalra, N., DiMartino, C., & Peng, A. X. (2011). *A big apple for educators: New York City's experiment with schoolwide performance bonuses.* Santa Mónica, California: RAND.

Martínez, J. F. (2012). Consequences of omitting the classroom in multilevel models of schooling: an illustration using the effects of opportunity to learn on reading achievement. *School Effectiveness and School Improvement, 23*(3), 305-326. doi:10.1080/09243453.2012.678864

Martínez, J. F., Borko, H., & Stecher, B. (2011). Measuring instructional practices in middle school science using classroom artifacts. *Journal for Research in Science Teaching, 41*(1), 38-67. doi: 10.1002/tea.20447

Mayer, D. (1999). Measuring instructional practice: can policy makers trust survey data? *Educational Evaluation and Policy Analysis, 21*(1), 29-45.

Mehrens, W. (1989). Combining evaluation data from multiple sources. En J. Millman y L. Darling-Hammond (Eds), *The new handbook of teacher evaluation: assessment of elementary and secondary school teachers* (pp. 322-336). Newbury Park, C. A.: Sage.

Mihaly, K., McCaffrey, D.F., Staiger, D.O., & Lockwood, J.R. (2013). *A Composite estimator of effective teaching*. Seattle, WA: Bill & Melinda Gates Foundation. Recuperado el 2 de enero de 2013 de http://www.metproject.org/downloads/MET_Composite_Estimator_of_Effective_Teaching_Research_Paper.pdf

Muthén, B. O., Huang, L. C., Jo, B., Khoo, S. T., Goff, G. N., Novak, J., & Shih, J. (1995). Opportunity-to-learn effects on achievement: analytical aspects. *Educational Evaluation and Policy Analysis, 17*, 371-403.

New Teacher Center (2009). *Validity and reliability of the North Carolina teacher working conditions survey*. Recuperado de http://www.ncteachingconditions.org/sites/default/files/attachments/validityandreliability.pdf

Peterson, K. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Education Research Journal*, *24*(2), 311-317.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: standardized observation can leverage capacity. *Educational Researcher*, *38*, 109–119. doi: 10.3102/0013189X09332374

Reynolds, M. (1999). Standards and professional practice: the TTA and initial teaching training. *British Journal of Educational Studies*, *47*(3), 247-260. doi: 10.1111/1467-8527.00117

Rowe, K. (2003). *The importance of techer quality as a key determinant of students' experiences and outcomes of schooling*. Trabajo presentado en la ACER Research Conference, Melbourne.

Rubin, D. B., Stuart, E.A., & Zanutto, E.L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29*(1), 103-116.

Santelices, V., & Taut, S. (2011). Convergent validity evidence regarding the validity of the Chilean standards-based teacher evaluation system. *Assessment in Education: Principles, Policy and Practice*, *18*(1), 73-93. doi: 10.1080/0969594X.2011.534948

Sartain, L., Stoelinga, S., & Brown, E. (2011). *Rethinking teacher evaluation in chicago: lessons from classroom observations, principal-teacher conferences, and district implementation, University of Chicago Consortium on School Research*. Recuperado el 8 de diciembre de 2012 http://ccsr.uchicago.edu/sites/default/files/publications/Teacher%20Eval%20Report%20FINAL.pdf

Schafer, W. D. (2003). A state perspective on multiple measures in school accountability. *Educational Measurement: Issues and Practice*, *22*, 27-31. doi: 10.1111/j.1745-3992.2003.tb00125.x

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: a review and resolution of the controversy. *Personnel Psychology, 24,* 419-434.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE 2010-4004). Washington, D. C.: National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, United States Department of Education.

Schweigh, J. (2012). *Cross-level measurement invariance in school and classroom environment variables*. [Manuscrito].

Sclafani, S., & Lim, E. (2008). *Rethinking human capital in education: Singapore as a model for teacher development*. Washington, D. C.: The Aspen Institute.

Shulman, L. (1998). Teacher portfolios: a theoretical activity. En N. Lyons (Ed.), *With portfolio in hand* (pp. 23-37). Nueva York: Teachers College Press.

Smith, P. S., & Taylor, M.J. (2010). *New tools for investigating the relationship between teacher content knowledge and student learning.* Trabajo presentado en la NARST Annual Conference, Philadelphia, PA.

Stecher, B., Le, V.N., Hamilton, L., Ryan, G., Robyn, A., & Lockwood, J. R. (2006). Using structured classroom vignettes to measure instructional practices in mathematics. *Educational Evaluation and Policy Analysis, 28*(2), 101-130.

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010). *Incorporating student performance measures into teacher evaluation systems*. Santa Monica, C. A.: The RAND Corporation.

Strunk, K., Weinstein, T., Makkonen, R., & Furedi, D. (2012). Lesson learned: Three lessons emerge from Los Angeles Unified School District's implementation of a new system for teacher evaluation, growth, and development. *Phi Delta Kappan*, 94(3), 47-51.

Taut, S., Santelices, M.V., & Stecher, B. (2012). Validation of a national teacher assessment and improvement system. *Educational Assessment*, *17*(4), 163-199. doi: 10.1080/10627197.2012.735913

Teacher Performance Assessment Consortium (2012). *About EDTPA*. Recuperado el 8 de diciembre de 2012 de http://edtpa.aacte.org/