Puntos de Referencia

*Edición online*
N° 433, julio 2016

# Medición de Valor Agregado de Profesores: Avances, desafíos y aplicaciones

**Fernando Ochoa**

## Resumen

En las últimas décadas las mediciones de valor agregado han captado la atención de académicos y *policy makers* principalmente por su bajo costo relativo y potencial efectividad como instrumento para medir indirectamente la calidad de un profesor. Sin embargo, han estado sujetas a críticas por el supuesto sesgo presente en ellas y posibles efectos indeseados sobre el comportamiento de profesores u otros agentes relevantes al ser aplicadas como política pública.

El objetivo de este trabajo es realizar una revisión de literatura sobre mediciones de valor agregado a profesores. En particular, se abordan tres temas principales: (i) La precisión y sesgo de las mediciones, (ii) efectos de mediano y largo plazo en alumnos de profesores de acuerdo a su nivel de valor agregado y (iii) los efectos de la aplicación de mediciones de valor agregado en la estructura de selección, evaluación e incentivos a profesores. Finalmente, dado que la reciente aprobación del proyecto de ley que cambia la Carrera Docente conllevará cambios en Sistema de Evaluación Docente, también se revisan las principales falencias de este y algunos factores a tomar en cuenta si se optara por incluir mediciones de valor agregado en él.

Las discusión sobre mediciones de valor agregado aún no está zanjada. Sin embargo, estudios recientes muestran que los niveles de sesgo en las mediciones de valor agregado son despreciables y que ser alumno de profesores de alto valor agregado tiene impactos positivos en el desempeño académico futuro, probabilidades de continuar estudios, nivel de ingresos una vez en el mercado laboral, entre otros. En cuanto a la aplicación de mediciones de valor agregado en estructuras de selección, evaluación e incentivos la evidencia es mixta, aunque se presentan efectos sistemáticamente superiores en países en desarrollo.

Hay relativo consenso en que las mediciones de valor agregado son un instrumento que puede agregar información relevante para identificar a profesores de alto o bajo rendimiento a un bajo costo, pero su efectividad es altamente sensible a que sean combinadas con medidas de calidad más amplias y un cuidadoso diseño de política que se adecue a cada sistema.

El Sistema de Evaluación Docente nacional ha sido valorado internacionalmente por ser pionero en medir de manera amplia la calidad de los profesores mediante diversos instrumentos. Con todo, las evaluaciones realizadas a este sistema y la experiencia comparada muestran que hay espacio para su mejora, principalmente eliminando componentes poco informativos. Se concluye que la inclusión de mediciones de valor agregado en el Sistema de Evaluación Docente puede ser una buena alternativa para mejorarlo y se entregan recomendaciones generales para que su aplicación sea exitosa.

**Fernando Ochoa**. Estudiante de Magister en Economía, Pontificia Universidad Católica de Chile. Licenciado en Economía, Universidad de Chile.

Agradezco los comentarios y sugerencias de Harald Beyer, Sylvia Eyzaguirre y Andrés Hernando. Cualquier error u omisión es de mi exclusiva responsabilidad.

## Introducción

La evaluación de la calidad de los profesores ha sido motivo de debate por décadas. Recientemente, el caso de las familias de nueve estudiantes que demandaron al Estado de California en Estados Unidos por rehusarse a despedir a los profesores de bajo desempeño remeció a ese país y volvió a encender el fuego cruzado entre académicos. Este caso captó la atención de destacados académicos del área, en particular Raj Chetty (Universidad de Harvard), John Friedman (Universidad de Brown) y Jonah Rockoff (Universidad de Columbia), quienes testificaron a favor de las familias argumentando que habían formas de identificar a los profesores de alto y bajo desempeño, en particular, empleando mediciones de valor agregado[1]. Al mismo tiempo, Jesse Rothstein (Universidad de California, Berkeley) testificó a favor del Estado de California, afirmando que las mediciones existentes, como las de valor agregado, no logran aislar completamente el efecto de los profesores, capturando en su desempeño una combinación de efectos externos al profesor y a la sala de clase, lo que produciría sesgos.

Si bien el fallo a favor de las familias fue histórico, estableciendo que la protección a los profesores por parte del Estado de California era inconstitucional, el debate sobre las evaluaciones a profesores sigue abierto. Las mediciones de valor agregado despiertan interés por su bajo costo y potencial utilidad para este fin, sin embargo, están sujetas a críticas en cuanto a su precisión y efectos al ser utilizadas para realizar política pública.

El presente trabajo intenta resumir la discusión en torno a las mediciones de valor agregado y su aplicación en la política pública. Por simplicidad, se agrupa el debate en tres discusiones: (i) Precisión y sesgo de las mediciones. (ii) Efectos de mediano y largo plazo en alumnos de los profesores de acuerdo su nivel de valor agregado. (iii) Efectos de la aplicación de mediciones de valor agregado en la estructura de evaluación e incentivos a profesores. Finalmente, se resume el caso de la Evaluación Docente en Chile y se concluye.

[1] El valor agregado del profesor es el efecto del profesor en el aprendizaje de sus alumnos, medido en términos de puntajes o notas en test o pruebas, en un período de tiempo controlando por las características de la sala de clase y de los alumnos.

## Qué entendemos por mediciones de valor agregado

La calidad de la educación es un concepto que es difícil de definir y, aún definido, complejo de medir. Son muchas las variables que pueden influir en la calidad de la educación, como, por ejemplo, la infraestructura, los métodos de enseñanza y el currículo académico. No obstante, hay relativo consenso y evidencia de que una de las variables preponderantes —especialmente para países en desarrollo— es la calidad de los profesores (Carneiro et al., 2015; Rivkin, 2005).

Es evidente que la calidad de un profesor no se puede reducir a una sola variable, no solo es relevante su nivel de conocimiento disciplinar y desempeño en la sala, sino también lo son elementos más abstractos como su habilidad para colaborar con sus colegas y comunidad, su capacidad de aportar al desarrollo integral del establecimiento educacional, entre otras cosas. Una forma de medir indirectamente la calidad de un profesor, que se ha popularizado y estudiado intensamente en la última década, son las mediciones de valor agregado (en adelante VA). Éstas buscan estimar el efecto promedio del profesor en algún indicador de avance educacional de sus alumnos, a menudo logros académicos específicos, en un periodo determinado, controlando por las características de la sala de clase y de cada alumno. La idea detrás de esta aproximación es que los profesores no son igualmente efectivos y que es importante intentar capturar la distribución de efectividad para definir políticas de compensación, contratación y apoyo.

El VA de un profesor se estima por medio de un modelo econométrico. Siguiendo a Hanushek (2010), la forma más general y simplificada[2] consiste en una función de producción del logro educativo[3], de la forma:

$$A_g = A_{g-1}\theta + \tau_j + S\varphi + X\gamma + \varepsilon$$

Donde $A_g$ representa alguna medición del logro académico del alumno[4] en el nivel g —medido ya sea por sus notas o por un test estandarizado—, $A_{g-1}$ es el logro académico en el nivel g–1, es un vector que contiene las características relevantes del colegio y los pares del alumno, y X un vector de factores familiares o del vecindario del niño. Tanto θ, φ y γ son parámetros desconocidos a estimar. El parámetro de interés para el análisis es $\tau_j$, un efecto fijo que representa el aprendizaje, en términos de puntaje en $A_g$, que es imputable solo al profesor j.

Los valores de los parámetros $\tau_j$, θ, φ y γ se estiman econométricamente. Los supuestos necesarios para obtener estimadores no sesgados no son triviales y aún cumpliéndose, al igual que en cualquier estimación econométrica, la duda persiste respecto de si capturan un efecto causal o una simple correlación estadística entre las variables consideradas. Por estos y otros motivos que se abordarán a continuación, las mediciones de VA y su utilización como una herramienta de política pública ha sido motivo de un extenso debate académico que en el último tiempo ha estado particularmente activo.

Las mediciones de VA son atractivas debido a que logran capturar indirectamente y a un costo relativamente bajo[5] la calidad de un profesor. No obstante, la discusión sobre sus posibles alcances y limitaciones aún está abierta. En adelante, se expone el estado del arte de la discusión en torno a la estimación y aplicación de mediciones del VA de los profesores. Los principales cuestionamientos y respectivas defensas, por simplicidad, son separadas en tres grandes grupos: (i) Críticas a la precisión y sesgo de las mediciones; (ii) efectos de largo plazo de profesores de alto VA sobre sus alumnos; (iii) implicancias de la aplicación de mediciones de VA en política pública y estructuras de incentivos sobre el comportamiento de los profesores, directores y otros agentes relevantes.

## 1. Precisión y sesgo de las mediciones de valor agregado

Debido a que la estimaciones de VA se obtienen, generalmente, con una regresión de Mínimos Cuadrados Ordinarios (MCO), el análisis descansa sobre los supuestos econométricos necesarios para conseguir una estimación no sesgada. En palabras simples, si se estima que el VA ($\tau_j$) de un profesor es una unidad, entonces los resultados de sus alumnos aumentan en promedio en una unidad, si los alumnos le fueran asignados al azar. Hanushek (2010) sostiene que la principal falencia de la mayoría de las estimaciones es la potencial endogeneidad en el modelo, cuya causa son principalmente las variables omitidas en la especificación de este[6]. En

[2] Existen versiones más complejas, en cuanto a la cantidad y tipos de controles, períodos de tiempo considerados y los pesos asignados a cada uno de ellos. Todas estas versiones presentan ventajas y desventajas. Sin embargo, no considero necesario abordar las versiones más complejas en este trabajo, dado que para el propósito de éste resulta suficiente la fórmula básica.

[3] Para una introducción amable a las funciones de producción de logro educacional o capital humano y las condiciones econométricas necesarias para obtener estimadores no sesgados ver Todd y Wolpin (2003).

[4] Subíndice omitido por simplicidad. Por la misma razón se omite el hecho de que en general se busca estimar el efecto del profesor sobre una materia en específico.

[5] Se requieren sólo datos administrativos sobre el desempeño del alumno y sus características individuales, familiares y de su vecindario. En cambio, mediciones más cualitativas exigen una evaluación caso a caso de cada profesor por parte de profesionales preparados para ello.

[6] El sesgo generado por la omisión de variables relevantes puede ejemplificarse con un modelo de regresión linear:

$$Y = \alpha + X\beta + W\delta + \varepsilon$$

Si este es el modelo real, que permite obtener el efecto causal de X sobre Y estimando β, pero estimamos el siguiente modelo omitiendo W:

$$Y = \alpha + X\beta - \varepsilon$$

El estimador MCO de β vendrá dado por:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Luego, reemplazando Y por la ecuación verdadera y calculando la esperanza condicional del estimador obtenemos:

$$E[\hat{\beta}|X] = \beta + (X'X)^{-1}X'W\delta$$

efecto, si el analista falla al no considerar o no tiene disponibles, por ejemplo, las variables asociadas a las decisiones de la escuela a la que asiste el niño (en particular, la distribución de los estudiantes entre los cursos y la forma en que los profesores son asignados a un grupo de estudiantes por el director de la escuela), se obtendrían estimadores y varianzas agregadas sesgadas.

Uno de los académicos más crítico de las mediciones de VA es Jesse Rothstein (Universidad de California, Berkeley). En su trabajo del año 2010 hace hincapié en que la asignación no aleatoria de alumnos a escuelas (procesos de selección) y profesores (*sorting*[7]) genera un sesgo en las mediciones de VA. Con datos para el Estado de Carolina del Norte, Rothstein estima un sesgo de hasta un 20% en la medición del VA para cada profesor. En particular, hace un test de falsificación[8] estimando el efecto de los futuros profesores en el puntaje de los alumnos —que debería tender a cero y ser no significativo si las estimaciones fueran no sesgadas y, por tanto, de efectos causales—, encontrando un impacto similar en magnitud y significancia estadística al de los profesores actuales y pasados. Rothstein argumenta que tales resultados serían evidencia de serios problemas de variables omitidas, indicando que los controles normalmente incluidos no logran capturar las decisiones de familias y colegios, de manera que parte de su efecto se atribuye al profesor de turno sesgando la estimación de su aporte al aprendizaje de los alumnos.

Baker et al. (2010) complementan la crítica sosteniendo que son muchos los factores que afectan los resultados potencialmente obtenidos por los alumnos en un test, cuyos resultados son luego erróneamente atribuidos —para bien o para mal— al profesor por medio de las mediciones de VA individuales. Por ejemplo, los resultados de un alumno pueden depender de los efectos de los profesores pasados y de los contemporáneos de otras asignaturas, atributos y recursos provistos por la escuela (tamaño de la sala de clases, tutorías especiales, acceso a material, motivación, etc.). Además, algunos factores externos asociados, por ejemplo, al barrio, la familia y los pares del alumno también serían relevantes y difíciles de capturar, aunque los factores más complejos e importantes de capturar serían las decisiones de familias (autoselección) y escuelas (selección y distribución de alumnos y profesores dentro de ellas), mencionadas anteriormente. Lo anterior implica castigar sistemáticamente en las mediciones de VA a los profesores que hacen clases en cursos con alumnos desaventajados intelectual o socioeconómicamente. Lo contrario ocurre con profesores que hacen clases a alumnos de más recursos, por ejemplo, los beneficios educativos de clases particulares, actividades extraprogramáticas o cursos de verano (frecuentes en EE. UU.) serían atribuidos, injustamente, a ellos.

Por otro lado, los defensores quizá más emblemáticos de las mediciones de VA son Raj Chetty, John Friedman y Jonah Rockoff (en adelante CFR)[9]. En un trabajo que ha sido muy influyente en la discusión, estos académicos evalúan el nivel de sesgo presente en las estimaciones más comunes de VA[10], utili-

Lo que significa que el estimador que obtenemos está sesgado en una magnitud y signo desconocidos, representado por el segundo término del lado derecho de la ecuación. Es trivial demostrar que si estimamos la ecuación considerando W obtenemos $E[\hat{\beta}|X] = \beta$, es decir, un estimador no sesgado.

[7] El *sorting* es la distribución de alumnos de un mismo nivel a distintos cursos basada en criterios no aleatorios, por ejemplo, tener un curso de "aventajados" y otro de "alumnos con problemas de aprendizaje".

[8] Un test de falsificación es una técnica frecuente en econometría utilizada para rechazar la idea de que una correlación encontrada en los datos es causal. Para ello se estima un modelo análogo al que se estudia, pero con relaciones que se saben no son causales. En caso de encontrar una correlación se acusa que la del modelo original es mecánica (no causal).

[9] Se consideran los más conocidos por su relevancia en el fallo judicial, que obligó al Estado de California a evaluar y despedir a los peores profesores.

[10] La única diferencia con las estimaciones convencionales (usadas por algunos estados) es que le dan mayor peso a las mediciones más recientes de VA. Las estimaciones convencionales, en general, dan igual peso al VA en el primer y último año trabajado. Esto no se ajusta a la realidad, dado que la calidad del profesor puede variar en el tiempo.

zando dos métodos: evaluación de la sensibilidad de la estimación a la inclusión/omisión de controles y un diseño cuasi-experimental[11] (CFR, 2014a). Para este estudio utilizaron datos administrativos para más de un millón de niños y sus respectivos profesores, siguiéndolos desde 4to grado (equivalente a cuarto básico) hasta su adultez. La base contiene información de más de 18 millones de test en lenguaje y matemáticas entre 1989 y 2009 que complementan con información de las características (ingreso del hogar, ahorros de pensión, edad de la madre cuando nace el hijo, entre otros) e impuestos pagados por las familias, así como información tributaria de los propios alumnos (impuestos pagados cuando son adultos, lo que permite estimar sus ingresos) desde 1996 a 2011.

En primer lugar, comparan los resultados de estimaciones de VA que controlan por las variables comúnmente utilizadas por los distritos escolares de EE.UU. (características del alumno, la clase, la escuela, entre otros) con los resultados de estimaciones que controlan por más variables, que se omiten normalmente (gracias a los datos administrativos de los que disponen cuentan con una batería de potenciales controles mucho más grande que la de estudios corrientes), principalmente información sobre los padres, como su ingreso y escolaridad. El resultado de esta comparación es un sesgo de máximo 2,6%. Debido a que se puede argumentar que esta corrección sería insuficiente para establecer causalidad (estimaciones no sesgadas)[12], los autores utilizan un cuasi-experimento para obtener efectos causales. El cuasi-experimento se basa en los cambios de profesores entre escuelas. La hipótesis es la siguiente: si un curso tenía un profesor de alto VA en un año y este profesor se cambia de colegio, al año siguiente el curso que deja debería obtener puntajes promedio menores después del cambio[13].

El ejercicio principal arroja una estimación del sesgo que no es significativamente distinta de cero a los niveles de confianza estadística usuales, mientras que la estimación que arroja el mayor sesgo estimado lo sitúa en torno del 9,1% de las estimaciones usuales de VA. Aún más, encuentran que incrementos (o disminuciones) en el VA de los profesores de una asignatura, aumentan (o disminuyen) significativamente los resultados del curso, sin encontrar una correlación con los puntajes pasados ni tampoco con otras asignaturas contemporáneas.

Una de las mayores contribuciones de CFR (2014a) es el estudio de los controles más relevantes con los que, de acuerdo a sus análisis, es necesario contar para obtener estimadores no sesgados usando el mecanismo tradicional. Al respecto, sostienen que al controlar por el puntaje obtenido en un test anterior de los alumnos se obtiene un sesgo estimado máximo de 5% y estadísticamente no diferente de cero. Kane et al. (2014) replican la metodología de los autores para *Los Ángeles Unified School District*, encontrando resultados consistentes con los de CFR. Además, los autores abordan directamente la crítica de Rothstein (2010), que apunta al sesgo por variables omitidas —en particular, selección y *sorting*—, concluyendo que el sesgo producido por estos motivos, luego de controlar por los puntajes de períodos anteriores, es prácticamente nulo. Es importante tener en cuenta que este resultado es válido para la base de datos que ellos analizan y aquellas en que ha sido posible

[11] En econometría se denomina cuasi-experimentos a cambios exógenos que se asemejan a una distribución aleatoria (experimento) y/o permiten identificar variaciones exógenas en variables endógenas.

[12] Esto se debe a que no es posible corregir por todas las posibles fuentes de endogeneidad (como el *sorting*).

[13] El supuesto de identificación sobre el que descansa la validez del diseño cuasi-experimental es que las altas frecuencias de movimiento de profesores dentro de la escuela no tienen correlación con las características de estudiantes o escuelas. Obviamente, este supuesto es altamente debatible en un contexto en el que las escuelas pueden competir por reclutar a los mejores profesores.

replicar su estudio, pero no se pueda extender a otros contextos o sistemas educativos sin más, pues diferencias en los niveles de selectividad y *sorting* entre ellos podría entregar resultados diferentes[14].

Si bien el trabajo de CFR (2014a) parece ser robusto y ha sido replicado por otros autores, está lejos de haber cerrado la discusión. Rothstein (2014) replica la estimación de los autores para una nueva base de datos de Carolina del Norte, encontrando una correlación entre los cambios de profesores y la preparación de los alumnos. Rothstein sostiene que los directores de colegios tienden a reemplazar a los profesores por otros de mayor VA en cursos que tienen una trayectoria creciente de desempeño por otros motivos, lo que invalidaría la estrategia de identificación basada en el cuasi-experimento utilizada por los autores. Luego, argumenta que las estimaciones de largo plazo no serían robustas y que sus resultados serían sensibles a los controles incluidos en el análisis.

Posteriormente CFR (2015) responden a las críticas de Rothstein. En lo medular, sostienen que la forma en que Rothstein imputa los datos perdidos (*missing data*) es incorrecta y que este error introduciría un sesgo en sus estimaciones que explicaría sus resultados. Por otro lado, la correlación que encontraría Rothstein entre el desempeño previo de los estudiantes y el VA del profesor sería un efecto mecánico, debido a que construye este último en términos del primero. Concluyen que el trabajo de Rothstein termina por confirmar los puntos hechos por ellos anteriormente (CFR, 2014a).

Además, en un reciente trabajo en desarrollo (*working paper)* de CFR (2016) se evalúa exhaustivamente la efectividad de utilizar los resultados pasados para predecir el sesgo en modelos de VA —el mismo que utiliza Rothstein en su trabajo de 2010 y 2014—, por medio de simulaciones de Monte Carlo. Estas simulaciones muestran que, a pesar de que sea un método intuitivo ("los profesores de hoy no deberían influir en los puntajes de ayer"), no es robusto para obtener información del sesgo de un modelo de VA a diferencia de otros contextos en que sí es efectivo (por ejemplo, estudios sobre el efecto de variaciones en el tamaño de la sala de clases sobre el rendimiento). Esto se debe a una serie de razones econométricas bastante técnicas, pero dicho de forma sencilla, para estimar el sesgo correctamente el modelo debe estar bien especificado y tener una base de datos grande, ambas condiciones son muy complejas de cumplir en el contexto de modelos de VA.

La discusión sobre el VA es bastante reciente, pero hasta el momento no han habido críticas sustantivas al trabajo de CFR ni una respuesta por parte de Rothstein.

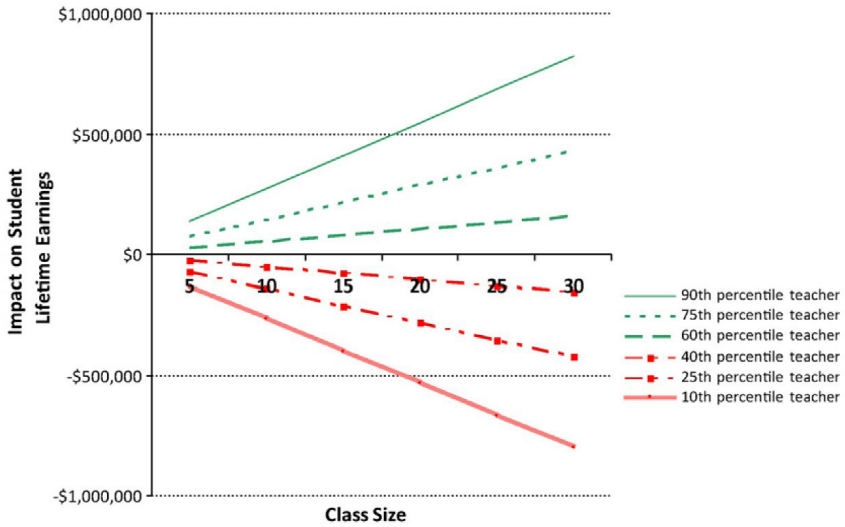
## 2. Efectos de mediano y largo plazo de profesores de alto valor agregado sobre sus alumnos

Una segunda arista del debate se ha centrado en si los alumnos de profesores de alto VA tienen mejores resultados sólo en los test empleados para la medición o si hay efectos en el desempeño académico futuro, probabilidades de continuar estudios, nivel de ingresos una vez en el mercado laboral, entre otros. Es decir, si los profesores que logran que sus alumnos tengan un desempeño superior a lo esperado en una determinada prueba también logran un aumento generalizado de sus oportunidades futuras.

Parte de la literatura ha tratado de traducir en retornos económicos futuros las ganancias de estar expuesto a profesores efectivos. El principal problema en esta discusión es que las estimaciones existentes asumen una relación causal del VA aportado por los profesores. Si por un momento asumimos que las

[14] Esto no es un problema particular asociado a las mediciones de VA, sino uno general en economía de la educación, salud, trabajo, entre otras áreas de estudio.

GRÁFICO 1: Impacto en los ingresos futuros de los estudiantes por tamaño de la sala de clases y calidad de los profesores (comparados con el profesor promedio)




Fuente: Hanushek (2011).

estimaciones de VA son correctas, hay una amplia literatura que se ha referido a la estrecha correlación entre los puntajes en test estandarizados e ingresos futuros. En particular, Rivkin et al. (2005) realizan una meticulosa estimación de los efectos de la calidad de los profesores y otros insumos escolares (tamaño de la sala de clases, material, etc.) en el rendimiento académico de los alumnos. Ellos concluyen que el factor determinante para el aprendizaje —en especial para los más desaventajados— es la calidad de los profesores y estiman que una reducción de diez alumnos por sala sería menos efectiva que aumentar en una desviación estándar la distribución de la calidad de los profesores.

Las conclusiones de Hanushek (2011) son aún más contundentes. Con estimaciones de valor agregado y un análisis de equilibrio parcial[15] estima el valor económico de la calidad de los profesores. Tal como se puede ver en el gráfico 1, Hanushek concluye que un profesor, cuyo VA es una desviación estándar sobre el promedio, genera ganancias de USD $ 400.000 en valor presente a estudiantes de una sala con 20 alumnos. Otra conclusión que obtiene Hanushek —quizá más discutible por ser de equilibrio parcial aplicado al crecimiento económico— es que, si se reemplaza entre el 5 y 8 por ciento de los peores profesores del país, EE.UU. quedaría en los primeros lugares de los rankings internacionales (por ejemplo en PISA) en las pruebas de matemática y ciencia, y con un efecto en crecimiento económico de USD $ 100 billones en valor presente (crecimiento del PIB).

Sin embargo, estas estimaciones están sujetas a críticas, porque, por una parte, asumen una estimación causal del VA y, por otra parte, existe evidencia que sostendría que las ganancias de los alumnos —expresadas en puntaje de test— de profesores de alto VA tienden a "desvanecerse en el tiempo". Por ejemplo, Kane y Staiger (2008) utilizan un experimento de asignación aleatoria de profesores[16] a 78 escuelas, 156 salas de clases y 3194 estudiantes. Su primera conclusión es que las mediciones de valor agregado no experimentales son bastante precisas, sin embargo, encuentran tanto en su diseño experimental como no experimental que el VA aportado por cada profesor se desvanece en el tiempo, decayendo a la mitad por cada año siguiente (observan hasta dos años). Poco se sabe de la razón de este

[15] Es decir, observando el efecto de los profesores se estima el efecto de un cambio en su calidad manteniendo todo lo demás constante (ceteris paribus). En equilibrio general (si se aplicara la política) no es claro si el impacto es menor, igual o mayor al estimado.

[16] Los experimentos aleatorios en ciencias sociales resultan ser un buen instrumento para identificar efectos causales de manera limpia. En este caso, al asignar aleatoriamente profesores y alumnos, se observa el efecto causal del profesor en el aprendizaje de los estudiantes resolviendo el problema de potenciales sesgos en las estimaciones.

desvanecimiento. Los autores sostienen que sería preocupante si se debe a que los alumnos olvidan lo que aprendieron o que el VA mide algo transitorio (como enseñar para la prueba), pero también reconocen que esto se puede deber a otras razones, como, por ejemplo, que lo enseñado no se evalúe después o que haya externalidades positivas —a través de efecto par— del alumno de un profesor de alto VA hacia sus compañeros, lo que haría indistinguible el efecto.

Lo anterior es congruente con lo que se ha encontrado en experimentos realizados en países en desarrollo. Banerjee et al. (2007) realizaron dos experimentos en India, donde forman a mujeres para enseñar en colegios a alumnos rezagados en matemática y lenguaje. Ellos encuentran efectos significativos, a saber, 0,28 desviaciones estándar en promedio con un impacto mayor en los alumnos más rezagados. Luego de un año, el efecto se mantiene significativo, pero 0,10 desviaciones estándar menor que el primer año. Una de las explicaciones para este fenómeno sería que los profesores (en este caso estudiantes egresados de enseñanza media capacitados para el programa) se centrarían en enseñar sólo el currículo, en vez de entregar una enseñanza más íntegra. Cabe destacar que este trabajo no se centra en mediciones de VA ni trabaja con profesores con educación superior completa, pero identifica cuidadosamente el efecto de desvanecimiento en el tiempo del efecto de las clases especiales en un contexto distinto.

Otro punto relevante respecto de los efectos de mediano y largo plazo tiene relación con el diseño de los test. Baker et al. (2010) afirman que incluso en los estados más avanzados en mediciones de VA, como New York y California, no se dispone de test completamente apropiados para este propósito. En primer lugar, desde el nivel preescolar hasta los niveles más avanzados de la enseñanza básica no existen test que logren capturar lo aprendido en un determinado periodo. Luego, los test existentes utilizados para estimar el VA por un profesor no están diseñados para eso, por lo que se generan problemas de atribución. Esta crítica atañe tanto al problema de estimación como a la utilidad en el largo plazo de las mediciones de VA. Por un lado, podemos estar atribuyendo (o restando) aportes a profesores en los distintos niveles y, por otro lado, se utilizan test diseñados para objetivos que no necesariamente son de evaluación de aprendizaje (los resultados pueden estar contaminados parcialmente por factores que no son cognitivos, por ejemplo, al ser decisivos para el éxito habilidades como la tolerancia a la frustración, concentración, entre otras). Esto significa que se podrían estar evaluando aptitudes que no necesariamente tendrán repercusión en estudios futuros, como por ejemplo el mercado laboral u otras dimensiones de la vida de los alumnos. Para evaluar correctamente el VA de un profesor los test deberían ser escalados verticalmente, es decir, evaluar lo que se va enseñando año a año en un tiempo continuo, usando el mismo instrumento.

Las implicancias prácticas de esta discusión son una pregunta abierta. Dependiendo del sistema de medición escogido, el VA podría ser un buen o mal instrumento, al tener mucho, poco o nada de sesgo debido al tipo de test utilizado. Precisamente por esto, la discusión anterior sobre el sesgo de las mediciones de VA es tan relevante. Si al estimar los niveles de sesgo en distintos contextos (con distintos sistemas de medición) se encuentran que son sistemáticamente irrelevantes, entonces esta discusión se resolvería indirectamente.

En un segundo trabajo, CFR (2014b) buscan estimar los efectos de los profesores de alto VA en los resultados de largo plazo de sus alumnos. Utilizando los datos mencionados anteriormente (para CFR, 2014a), los autores son capaces de seguir el historial de más de un millón de alumnos desde cuarto

grado hasta la adultez temprana. Los autores estudian la relación entre los salarios de largo plazo, la probabilidad de matricularse en la universidad y el embarazo adolescente de los alumnos con los resultados de sus profesores en sus dos estimaciones de VA previas (CFR, 2014a). El trabajo concluye que, haber sido alumno de profesores de mayor VA se traduce en incrementos significativos en la probabilidad de ingresar a la universidad, así como de la calidad de la universidad a la que se asiste[17]. También aumentan las trayectorias de ingreso durante la adultez temprana. En efecto, un incremento en 1 desviación estándar de la calidad de los profesores (VA) en un solo grado aumentaría en 1,3% el ingreso a los 28 años. Si se considera esa tasa de aumento constante en los años venideros (supuesto conservador según los autores), los alumnos ganarían en promedio USD $ 39.000 extras en su ciclo de vida, por el solo efecto de haber tenido un profesor que pertenece al 18% superior en términos de VA[18]. Haber sido alumno de profesores con mayor valor agregado correlaciona también negativamente con la probabilidad de embarazo adolescente y positivamente con la calidad del barrio en que se vive y la participación en los planes de pensiones.

Ante esta evidencia los autores concluyen que las preocupaciones por el decaimiento de los aportes de los profesores de alto VA en el tiempo no se sostienen. De hecho, confirman una alta persistencia del efecto del profesor, aunque sostienen que existiría un fenómeno de reaparición en forma de resultados de largo plazo, que seguiría los patrones encontrados en intervenciones de educación temprana (Heckman et al., 2010).

[17] Medida como el ingreso promedio de sus egresados, asumiendo que universidades de mayor calidad tienden a producir egresados que son mejor remunerados. Por supuesto, esta medida no está exenta de problemas de selección y *sorting*.

[18] Se supone una distribución normal de VA.

## 3. Implicancias del uso de mediciones de valor agregados sobre estructuras de incentivos a profesores y políticas públicas asociadas

Esta sección aborda la aplicación de mediciones de VA como insumos para sistemas de selección e incentivos a profesores como política pública. Los sistemas de incentivos a profesores tienen dos objetivos principales: primero, buscar que los profesores tengan motivaciones para "esforzarse" en el cumplimiento de su labor, ya sea en términos de cantidad (cumplimiento de las horas dedicadas) o calidad (utilizar los mejores métodos de enseñanza posibles); segundo, atraer a los mejores individuos —en términos de productividad y/o vocación— a la profesión docente.

Para lograr los objetivos enunciados en el párrafo precedente se asocian parte de las remuneraciones o regalías que recibe un profesor con su resultado en alguna medición de la calidad de su trabajo. De esta forma, se pretende premiar a los profesores que consiguen mejores resultados de aprendizaje en la sala de clases, lo que induciría a estos a esforzarse por conseguir dichos resultados. En este sentido, en algunos sistemas se utilizan mediciones de VA como una medida indirecta de la calidad del trabajo docente. Sin embargo, hay una serie de *trade-offs* y problemas de diseño asociados, que despiertan el debate de académicos, *policy makers* y políticos.

Las críticas a la aplicación de estructuras de incentivos basadas en mediciones de VA tienen que ver con posibles efectos no deseados relacionados al comportamiento de profesores y directores, además de las derivadas de las críticas abordadas anteriormente. Baker et al. (2010) identifican dentro de los principales efectos no deseados: (i) Desincentivos a trabajar con los estudiantes más desaventajados dentro del colegio y/o sala de clases. Al no poder controlarse la estimación de VA por

la habilidad y contexto emocional —por ejemplo, efectos de la dinámica familiar del alumno—, no habría incentivos a dedicar más esfuerzo a estos niños, ya que tendrían un "menor retorno esperado". (ii) Estrechamiento del currículo, debido a que los profesores y escuelas tendrían incentivos para enfocarse en enseñar los contenidos evaluados en las pruebas que cuentan para las estimaciones de VA y sus respectivos métodos de respuesta (generalmente ítems de selección múltiple), dejando de lado aspectos relevantes para una formación integral y una acumulación de conocimiento más amplia. (iii) Menos colaboración entre profesores. Las mediciones de VA son aplicadas generalmente a nivel individual, aunque se pueden aplicar a cualquier nivel de agregación como cursos, niveles u otros. A pesar de que los incentivos basados en VA no promueven la competencia entre colegas, al importar únicamente el desempeño individual no incentiva la colaboración con otros profesores de la escuela, aunque tampoco la inhibe. Por otro lado, la opción de aplicar mediciones de VA a nivel grupal generaría problemas de *free riding*[19], afectando negativamente el impacto de la política en cuestión. Escoger el nivel de agrupación óptimo no es trivial y un mal diseño puede generar resultados nulos e incluso negativos. (iv) Desmoralización. La presión por obtener resultados en un test puede frustrar a buenos profesores (que tienen interés en desarrollar otras áreas del conocimiento también) y en especial a los que tienen alumnos vulnerables, debido a que por factores externos a la sala de clases los resultados en dichos test pueden ser sistemáticamente mediocres.

El principal problema para testear empíricamente si los efectos teóricos antes enunciados influyen en los resultados educacionales, cuando se establecen sistemas de incentivos basados en mediciones de VA, es que no basta con observar las políticas efectivamente implementadas en distintos sistemas educativos, toda vez que el diseño de los mismos puede responder a características propias no observables del ambiente en que se implementan, como por ejemplo la falta de profesores o de otros recursos. Técnicamente, este problema genera dificultades para obtener un contrafactual válido. Por esto, la literatura en los últimos años se ha apoyado en diseños experimentales para responder las preguntas en torno a los efectos de las estructuras de incentivos a profesores y, en los casos particulares que nos interesan, de las mediciones de VA.

Un estudio experimental que aborda directamente esta problemática es el de Glewwe et al. (2010). Ellos examinan los efectos de un experimento aplicado en Kenya que premia[20] a los profesores en base a resultados de sus alumnos en un test nacional. Simultáneamente aplican otro tipo de pruebas —preguntas abiertas en vez de alternativas, por ejemplo— a los alumnos, por las cuales no hay premios asociados. Los investigadores encuentran que hay un impacto positivo sólo en los puntajes del test que era premiado y un aumento de las tutorías de preparación para el examen para el cual había incentivos, indicando que los profesores educarían "para la prueba", en particular habilidades para la resolución de preguntas de alternativas. Además, encuentran que las mejoras son mayores en las materias que son intensivas en memorización, teniendo ganancias principalmente en geografía, historia y religión. En matemática y ciencias natu-

[19] En ciencias sociales se llama *free rider* o "polizón" a un individuo que consume más de lo equitativo o no enfrenta los costos de producción de un bien público (o comunitario), pero aun así goza de sus beneficios. En nuestro caso, la lógica es la siguiente: frente a un sistema de incentivos a profesores por el desempeño grupal, cada uno de ellos decide su nivel de esfuerzo óptimo estimando los costos y beneficios (económicos, sociales, personales, etc.). Una posible decisión es esforzarse poco debido a que el impacto de su actuar es marginal (representa poco del resultado final) y puede beneficiarse del esfuerzo de los demás (recibir igual el premio). En un contexto en que esa decisión es la óptima para todos o una gran parte del grupo definido para el incentivo, se obtiene un nivel de esfuerzo promedio subóptimo. Debido a esto, la elección del "tamaño del grupo" (colegio, nivel, asignatura, curso u otro) que estará sujeto al incentivo afecta la efectividad de la política en cuestión.

[20] Los incentivos no son pecuniarios por razones culturales.

TABLA 1: Resumen de evaluaciones experimentales que utilizan VA de profesores y que han tenido más impacto en la literatura reciente

| Estudio | Lugar (Nombre del Programa) | Tipo de Incentivo | Premio máximo | Efecto en Matemática (D.E.) | Efecto en Lenguaje/ Lectura (D.E.) |
|---|---|---|---|---|---|
| Dee & Wyckoff (2013) | Washington DC, USA (IMPACT) | Premios individuales basados en observaciones, involucramiento con la escuela y mediciones de VA por pruebas. | USD$27.000 por año, permanentemente | 0.24 | |
| Glazerman & Seifullah (2012) | Chicago IL, USA (Teacher Advancement Program) | Premios individuales basados en observaciones, involucramiento con la escuela y mediciones de VA por pruebas. | Bono de USD$6.400, una vez. | -0.03 | 0.01 |
| Imberman & Lovenheim (A) (2015) | Houston, TX, USA | Premios por grupo a nivel de asignatura basados en mediciones de VA. | Bono de USD$7.700, una vez. | 0.10* | 0.03 |
| Lavy (2009) | Israel | Premios individuales basados en puntajes en pruebas. | Bono de USD$7.500, una vez. | (B) | |
| Muralidharan & Sundararaman (2011) | | | | | |
| | Andhra, Pradesh e India | Premios individuales o grupales a nivel escolar por ganancias en puntajes en pruebas (VA). | USD$11 por punto porcentual ganado en el promedio. | 0.28* | 0.17* |
| Springer et al. (2012) | Nashville, TN, USA (POINT Experiment) | Premios individuales basados en VA. | Bono por USD$15.000, una vez. | 0.05 | -- |

Notas: Si sólo se muestra una estimación, entonces representa un promedio de múltiples mediciones. A.- El estudio no identifica los impactos directos, pero lo hace indirectamente a través del efecto de cambiar el impacto del profesor sobre la probabilidad de recibir el premio. B.- No provee información para calcular el tamaño del efecto pero encuentra impactos positivos en ambos. * Resultado estadísticamente significativo. D.E.=Desviación Estándar.
Fuente: Extracto "Figura 1" de Imberman (2015).

rales las ganancias fueron más modestas y en otras asignaturas como lenguaje no obtuvieron efectos estadísticamente significativos. Es importante notar que los autores no concluyen que esto sea negativo, ya que podría estar compensando las diferencias con los alumnos de más recursos y un análisis de equilibrio parcial sería incompleto.

Imberman (2015) resume las evaluaciones experimentales que han tenido más impacto en la literatura reciente. Si bien su trabajo se enfoca en los incentivos financieros en general a profesores, en esta sección discutiremos solo los que utilizan mediciones de VA en su diseño. La tabla 1 es un extracto de la tabla resumen de su trabajo.

Luego de analizar la evidencia para EE.UU. y países en desarrollo, Imberman (2015) concluye que es mixta. Para los países en desarrollo, los incentivos parecen ser exitosos, pero para países desarrollados como EE.UU. e Israel los efectos no son claros, aunque con una tendencia a estimar resultados positivos. Sin embargo, para ambos casos parece comprobarse la tesis de que las mejoras en el rendimiento de los estudiantes se concentran en los test a los que están asociados los incentivos, pero prácticamente no se presentan mejoras en otras medidas de aprendizaje. No obstante, esto no significa que en ausencia de estos test los alumnos aprenderían otras cosas. Frente a la ausencia de un

contrafactual válido, este punto debe ser considerado con cuidado.

A pesar de lo anterior, Imberman (2015) sostiene que testear otras tesis se vuelve complejo. En primer lugar, los estudios se concentran en EE.UU. con muy poca evidencia rigurosa para países en desarrollo y prácticamente inexistente para países Latinoamericanos. Además, la mayoría de los estudios asocia los incentivos a la superación de umbrales o metas de mejora y estos esquemas de incentivos son precisamente los que tienden a tener menores efectos. Esto se debe al desincentivo al esfuerzo para profesores cuya condición inicial los deja muy lejos de la meta. Por el contrario, diseños que premian las ganancias marginales tienden a tener mayores efectos.

Si bien el sistema de metas es atractivo para los estados en términos de manejo del presupuesto, como se observa pueden ser poco efectivos, teniendo sólo efectos en los profesores o colegios que ya se encuentran "cercanos a la meta".

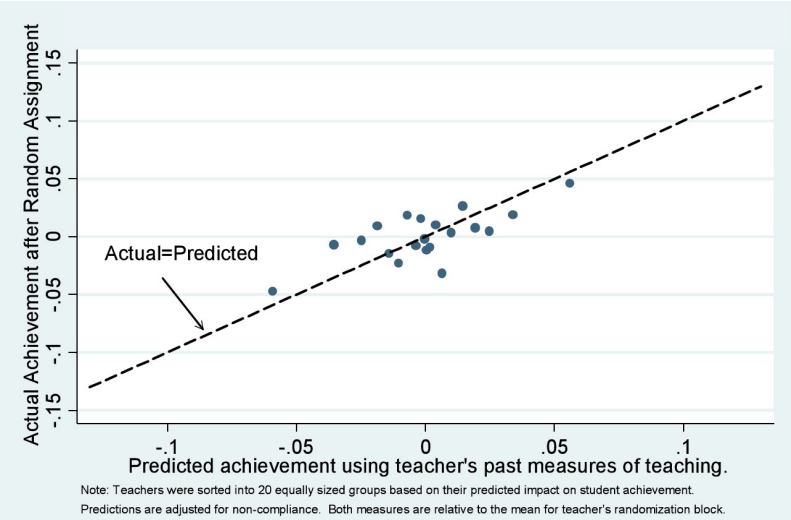
Al considerar la existencia de diferencias entre incentivos individuales frente a grupales, la evidencia puede engañar, debido a la poca comparabilidad de los estudios y sus diseños. En efecto, son prácticamente inexistentes los análisis que comparan ambos tipos de incentivos en un mismo contexto, sin embargo, los incentivos para grupos grandes (como la escuela) parecen ser inefectivos, siendo recomendable grupos de menor tamaño (por curso o asignatura) que eviten problemas de *free riding*.

Por último, poco se ha estudiado sobre los efectos de metas múltiples frente a particulares (como sólo premiar el VA), aunque la evidencia sugeriría que, mientras más amplias son las metas y mediciones, mayores son los efectos.

Un punto relevante a destacar en esta discusión es que incluso los académicos más críticos de las mediciones de VA (ver, por ejemplo, Baker et al., 2010) no proponen su total eliminación, sino tratar las mediciones con especial cautela, a saber, no darles una ponderación mayoritaria y combinarlas con otros tipos de evaluaciones de desempeño en el aula, entre pares y grupos. Al mismo tiempo, los defensores de la medición de VA no proponen que sea la única herramienta utilizada para establecer sistemas de incentivos, sino que la presentan como una herramienta útil, pero incompleta para capturar la calidad de la labor docente (Chetty, Friedman y Rockoff, 2015a, b).

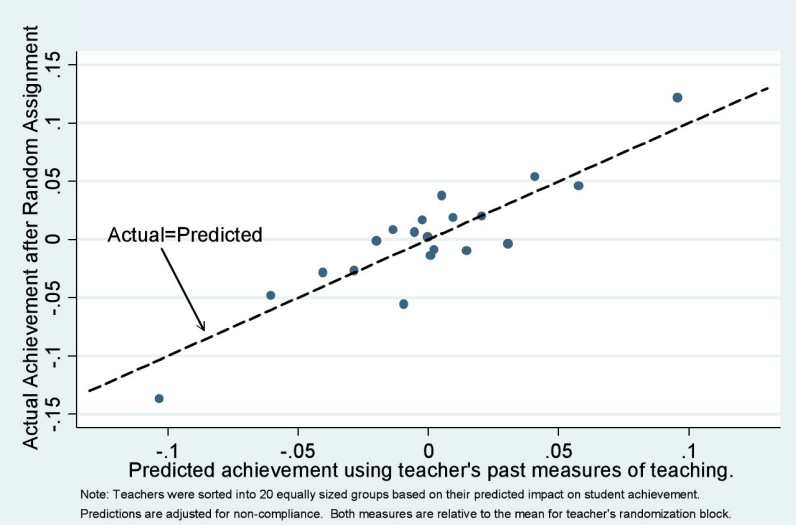
En línea con el punto anterior, un trabajo interesante es el proyecto *Measures of Efective Teaching* (MET) de la Bill & Melinda Gates Foundation, un trabajo en conjunto con académicos, profesores y organizaciones educativas, que investigaron formas de evaluar el desempeño en aula de profesores. En relación con las mediciones de VA, Kane et al. (2013) intentan hacerse cargo de las críticas en cuanto al sesgo y la incertidumbre frente a la combinación de medidas por medio de un diseño experimental. Recolectaron información durante tres años (2008-2011), tanto de resultados de escuelas en test estatales de EE.UU. (equivalentes al SIMCE), como del desempeño de los profesores en otras medidas de efectividad, evaluación de las clases (videos) y encuestas a los estudiantes sobre su percepción del profesor y el ambiente en la sala de clases. En una primera etapa (2008-2009) obtienen los resultados de las medidas y desempeño de los estudiantes en un contexto no experimental, luego (2010-2011) generan un experimento de asignación aleatoria de alumnos a profesores para así comparar ambos resultados y evaluar el potencial sesgo. Utilizan distintas especificaciones de VA, siendo la principal para el estudio la "medición compuesta", que además de controlar por el logro académico pasado del alumno lo hace por el puntaje del profesor en la evaluación de los videos de sus clases y las encuestas hechas a sus alumnos. Los resultados muestran que los efectos causales

GRÁFICO 2: Logro académico estimado y efectivo de las salas de clase aleatorizadas (Matemática)




Fuente: Extracto "Figura 1" de Kane et al. (2013).

GRÁFICO 3: Logro académico estimado y efectivo de las salas de clase aleatorizadas (Inglés y literatura)




Fuente: Extracto "Figura 1" de Kane et al. (2013)

(experimentales) son en promedio equivalentes a los estimados en un contexto no experimental (donde existe *sorting*). Una muestra gráfica de esto se puede apreciar en los gráficos 2 y 3. La principal limitación de este estudio es que el experimento se realizó a nivel de la escuela, no entre escuelas —por la evidente dificultad de asignar aleatoriamente profesores y alumnos a distintas escuelas—, lo que hace legítimo el cuestionamiento a la validez externa de los resultados. Además, se encuentra que los profesores de alto VA compuesto no sólo incrementan los resultados de sus alumnos en los test estatales, sino que también en otros test más demandantes cognitivamente, tanto en matemática como lenguaje (aunque el efecto es alrededor de un tercio menor).

Lo anterior es especialmente relevante porque otros trabajos del MET (Mihaly et al., 2013) muestran que al utilizar medidas de VA comunes (no compuestas) la correlación con el impacto en otros test (los no evaluados) es menor. De esta forma, medidas compuestas de evaluación a profesores permitirían una mejor identificación de los profesores efectivos (y no efectivos).

Kane y Staiger (2012) se preguntan cuál es la mejor forma de combinar las distintas evaluaciones a profesores para obtener un índice de calidad que permita dar *feedback* a los profesores. Los resultados de la investigación indican que la ponderación para las mediciones basadas en test estandarizados (VA) debe encontrarse entre un 33 y 50 por ciento. En ese rango se minimiza la volatilidad del índice entre años[21].

[21] Si bien se gana confianza combinando los diferentes instrumentos y se mejora la predicción del desempeño de los alumnos en test más complejos cognitivamente, se pierde predictibilidad respecto de los resultados en los test estandarizados estatales.

Independiente de lo relativamente incompleta de la literatura en este campo, tanto críticos como promotores de las mediciones de valor agregado concuerdan en que la aplicación a gran escala de éstas con diseños de política pública poco cuidadosos pueden generar efectos indeseados, pues pueden incentivar "enseñar para la prueba", hacer trampa en mediciones, contrarrestar los incentivos a colaborar con colegas, disminuir el esfuerzo en la enseñanza de tópicos no evaluados (como las artes), desincentivar a buenos profesores a trabajar en escuelas con niños más necesitados, entre otros. Al mismo tiempo, ambos grupos de investigadores resaltan la importancia de avanzar en la medición de la calidad de profesores por medio de evaluaciones más completas e integrales, tanto a nivel individual como grupal, siendo la medición de VA una aproximación indirecta de la calidad del profesor, que sería útil en conjunto con otras que la complementaran. Lamentablemente, hay menos consenso aún, tanto en el área de la educación como la economía, en cómo medir la calidad del profesor de forma directa y/o más amplia.

## La evaluación docente en Chile

Desde 2003, Chile comenzó a implementar el Sistema de Evaluación del Desempeño Profesional Docente con el fin de evaluar a los profesores del sector público a nivel nacional. Con esto se convirtió en uno de los países pioneros en la región en aplicar este tipo de evaluaciones[22]. No obstante, aún hay una serie de preguntas sin respuestas y desafíos para la mejora del sistema, especialmente si se optara por incluir mediciones de VA en la Evaluación Docente.

Actualmente, los profesores deben someterse a la Evaluación Docente cada cuatro años, siendo calificados según su nivel de desempeño como Destacado, Competente, Básico o Insatisfactorio. Desde 2011, en caso de quedar en el tramo Básico o Insatisfactorio el docente es evaluado nuevamente al año siguiente. Si es evaluado como insatisfactorio en dos evaluaciones consecutivas o como básico en tres evaluaciones consecutivas (o en forma alternada con desempeño básico o insatisfactorio), el docente es desvinculado de la dotación docente.

La evidencia para la evaluación se recoge por medio de cuatro instrumentos que deben guardar relación con lo estipulado por el Marco para la Buena Enseñanza y ser aprobados por el Centro de Perfeccionamiento, Experimentación e Investigaciones Pedagógicas (CPEIP)[23]. Estos son:

- Autoevaluación: El docente reflexiona en torno a una pauta generada por el CPEIP sobre su desempeño.
- Entrevista por un evaluador par: Es realizada por un docente par previamente capacitado. Se efectúa en base a una pauta que contiene preguntas sobre la propia labor docente y el entorno en el cual trabaja.
- Informe de referencia de terceros: Pauta que debe ser completada por el director y jefe de la Unidad Técnico Profesional (UTP) del establecimiento donde trabaja el docente.
- Portafolio de desempeño pedagógico: Consiste en la presentación de productos escritos y la grabación de una clase de cuarenta minutos de duración. La evaluación es realizada por docentes previamente capacitados del mismo nivel y asignatura.

En cada instrumento se evalúa al profesor como Destacado, Competente, Básico o Insatisfactorio. Cada instrumento posee una ponderación distinta para la

[22] Para una minuciosa revisión de la historia del Sistema de Evaluación del Desempeño Profesional Docente revisar Manzi et al. (2011).

[23] Ver www.docentemas.cl.

evaluación final. En caso de ser la primera evaluación en cuatro años, las ponderaciones son: 10, 20, 10 y 60 por ciento, respectivamente. En caso de existir una evaluación previa en nivel insatisfactorio la ponderación cambia a 5, 10, 5 y 80 por ciento[24].

Si bien la Evaluación Docente en Chile ha sido destacada a nivel regional e incluso mundial por los esfuerzos que hace para integrar distintas medidas cualitativas del desempeño docente, aún es relativamente escasa la evidencia sobre el vínculo de ésta con el desempeño de los estudiantes. La principal limitación que tienen los estudios existentes es que se basan en correlaciones estadísticas en vez de efectos causales.

El primer trabajo que trató de evaluar la relación entre la Evaluación Docente y el desempeño académico de los alumnos, es decir, si los docentes que están siendo evaluados en los tramos más altos al mismo tiempo generan los mejores resultados en sus alumnos, fue el de Bravo et al. (2008). Los autores estiman una función de producción que explica los resultados en la prueba SIMCE a partir de vectores de características del estudiante, la familia y el docente, encontrándose en este último los resultados de la Evaluación Docente. Debido a que no disponen de datos de panel para controlar por *sorting* (endogeneidad), agregan como control el puntaje previo en SIMCE del establecimiento. Si bien es mejor que no controlar, no es un método suficiente para establecer causalidad, por lo que sólo nos encontramos con correlaciones estadísticas. Los autores encuentran una relación positiva y estadísticamente significativa entre la Evaluación Docente y el desempeño de los estudiantes en SIMCE. En promedio, la diferencia entre un profesor Destacado y uno Insatisfactorio equivaldría al de tener padres con estudios universitarios o con enseñanza media (el efecto es mayor en establecimientos de menores recursos).

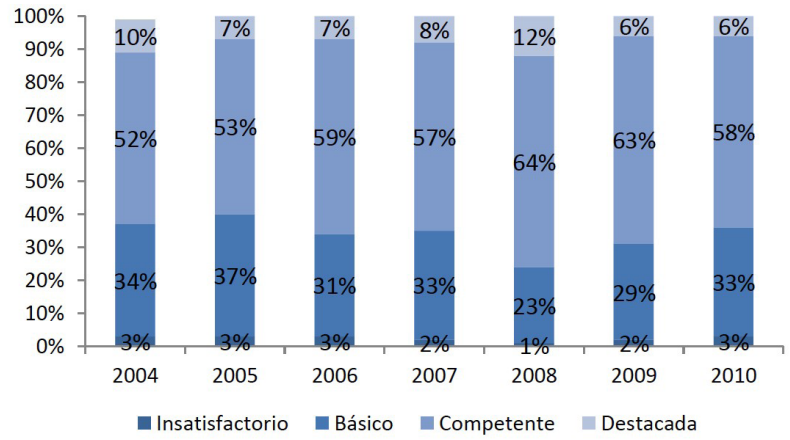
Un estudio posterior fue realizado en el marco de la auditoría al sistema de Evaluación Docente, cuya parte cualitativa fue encargada a la OCDE y la cuantitativa al PNUD. El estudio cuantitativo (Alvarado et al., 2012) utiliza distintas metodologías, la mayoría semejantes a la de Bravo et al. (2008), siendo capaz de encontrar solamente correlaciones estadísticas en vez de efectos causales. Los autores encuentran que hay una correlación positiva entre el desempeño de los docentes en la Evaluación Docente y el de los estudiantes; un punto más en la Evaluación Docente entre 1º y 4º básico se relaciona con 15,8 puntos más en la prueba SIMCE de matemática y 9,5 en la de lectura.

Lo interesante del trabajo de Alvarado et al. (2012) es el minucioso análisis de las diferencias en impacto entre los tramos en que son calificados los profesores y de la correlación de cada instrumento con los resultados académicos. Ellos encuentran que las diferencias son estadísticamente significativas al comparar profesores del tramo Competente o Destacado con el Insatisfactorio, no así al comparar los del tramo Básico con este último. Por otro lado, la correlación de los distintos instrumentos en el resultado académico de los alumnos es bastante diferente. Mientras un punto adicional en el portafolio se asocia con 8 puntos en matemática y 6 en lectura, el informe de referencia de terceros lo hace con 4,2 y 2,2 puntos más, respectivamente, y la entrevista por un evaluador par se vincula con 2,6 y 1 punto, respectivamente. Por último, la autoevaluación no muestra relación significativa con ninguna de las pruebas SIMCE. El trabajo concluye que el portafolio posee una distribución que se aproxima a una normal, mientras los otros tres instrumentos se encuentran sesgados hacia puntajes superiores. Surge la interrogante de si las evaluaciones son

[24] Para una revisión más detallada de la estructura de la evaluación docente, revisar Alvarado et al. (2012). Informe encargado por el MINEDUC al PNUD para una auditoría cuantitativa de la Evaluación Docente en Chile.

contradictorias entre sí, al haber profesores Destacados en ciertos instrumentos que al mismo tiempo son Insatisfactorios en otros. Esto puede ser cierto, si los instrumentos apuntan a medir la calidad docente en general, y falso, si cada uno midiese áreas independientes entre sí. Ahora bien, aunque así fuese, es cuestionable que existan diferencias tan marcadas en los resultados de los diferentes instrumentos. Más grave aún, el ordenamiento de los profesores en los cuartiles es sensible a la inclusión de instrumentos como la autoevaluación, debido a que un instrumento que no posee correlación estadística con el desempeño de los alumnos podría hacer pasar a un profesor de un nivel a otro.

GRÁFICO 4: Distribución de resultados en la Evaluación Docente por año




Fuente: Extracto "Figura 2" de Alvarado et al. (2012).

Un factor relevante para la creación de estructuras de incentivos es tener una dispersión de resultados en la evaluación docente para identificar de mejor manera a los distintos tipos de docentes, con el fin de seleccionarlos, premiarlos e incentivar su mejora constante. Hay un relativo consenso de que los directores de establecimientos pueden tomar mejores decisiones sobre los docentes en los extremos de las distribuciones, especialmente si se combinan mediciones cuantitativas (como de VA) con cualitativas (Jacob y Lefgren, 2008).

Como se puede aprecia en el gráfico 4, la distribución de los docentes del sector municipal está muy concentrada en el nivel Competente, con pocos profesores en los extremos. Esto limita tanto la aplicación de políticas públicas focalizadas en los distintos tipos de profesores (no se pueden identificar) y limita la capacidad de acción de los Directores de establecimientos.

En resumen, no existen estudios que muestren evidencia causal del vínculo entre la evaluación docente nacional y el desempeño académico de nuestros estudiantes. Es más, la evidencia existente da cuenta de ciertas deficiencias en el diseño del sistema. No obstante, crecientemente se está vinculando los resultados de la evaluación docente con remuneraciones, como es el caso de la nueva ley de política docente. Independiente de la aplicación de mediciones de VA, urge revisar la batería de instrumentos y su ponderación en la evaluación final, modificando o alterando los instrumentos que no entreguen información relevante y sesguen los resultados injustamente.

En caso de optar por la aplicación de mediciones de VA, la mayor limitación de la institucionalidad es la inexistencia de pruebas estandarizadas periódicas que permitan la estimación, además de evaluar si las pruebas existentes son apropiadas para este fin (Martínez, 2012)[25]. En cualquier caso, es necesario un cuidadoso diseño de la política de evaluación y los incentivos asociados, idealmente testeando previamente niveles de sesgo e impacto de los programas.

[25] Si bien el autor está en contra de agregar mediciones de VA en la Evaluación Docente en Chile, una revisión más minuciosa de este punto se puede encontrar en Martínez (2011).

## Conclusión

A pesar de ser un área de investigación bastante activa en el último tiempo y de los innegables progresos que se han realizado en la misma, la discusión en torno a la forma de realizar mediciones apropiadas del valor agregado de un profesor y aplicarlas como política pública está aún lejos de cerrarse. Si bien aún persisten interrogantes respecto al nivel de sesgo que incluyen las mediciones tradicionales y al verdadero impacto de largo plazo de la calidad de los profesores en el desempeño y la calidad de vida de sus alumnos medida, indirectamente, en términos de VA, la investigación reciente apoya la idea de que el sesgo obtenido por estimaciones no experimentales es relativamente bajo o inexistente. También, aunque con un menor número de investigaciones, se sostiene la idea de que la exposición de un alumno a profesores de mayor VA tendría un impacto de largo plazo. Finalmente, es escasa la investigación relacionada con la aplicación de mediciones de VA a las estructuras de selección e incentivos a profesores, aunque la evidencia existente sugiere que la combinación de estas medidas con otras más cualitativas permitiría identificar, premiar y apoyar a los distintos tipos de profesores de forma efectiva.

En los últimos años variados sistemas —principalmente estados de EE.UU.— han optado por incluir tales mediciones y estructuras de incentivos asociadas a ellas con resultados que aún no han sido analizados exhaustivamente.

Un punto a resaltar es que ninguno de los trabajos principales estudiados en este análisis plantea eliminar las evaluaciones de VA, sino complementarlas con otros instrumentos de evaluación individual y grupal más amplios. Los desacuerdos entre los autores se concentran en el nivel de confianza y relevancia que se puede conceder razonablemente a mediciones de VA en esta ecuación. Al mismo tiempo, se hace hincapié en la complejidad del sistema de incentivos asociado a la evaluación individual del VA como política pública, ya que puede llevar a comportamientos indeseados por parte de los profesores y direcciones de escuelas. Los *policy makers* deben estar conscientes que se requiere un análisis exhaustivo de la herramienta y del diseño de política a fin de evitar tales inconvenientes al momento de su implementación.

Es importante destacar que la evidencia analizada corresponde, en su mayoría, a estudios realizados en EE.UU., lo que se debe a la disponibilidad de datos de la calidad necesaria para realizar estimaciones precisas de los efectos considerados. Una preocupación extra, entonces, es la validez externa de estos estudios. En efecto, dado que el sistema estadounidense difiere de manera importante en sus características y diseño del sistema educacional chileno en aspectos como la selectividad o la segregación, es razonable esperar que análisis de VA implementados en Chile u otros países puedan disminuir o exacerbar los problemas asociados a las variables omitidas en las estimaciones. Esta posibilidad, que debe ser tenida en cuenta al momento de implementación de la política, sólo puede ser analizada y cuantificada con estudios empíricos que cuenten con el diseño adecuado para identificar exitosamente los efectos de la medición y su asociación con incentivos para los docentes y escuelas.

Si bien parece deseable optar por la aplicación de mediciones de VA en la batería de instrumentos utilizados para la medición de la calidad de los docentes, ésta política pública debe implementarse con cautela, acompañada de un análisis de la cantidad de información necesaria para reducir tanto como sea posible la cantidad de variables relevantes omitidas en el análisis, dependiendo esto crucialmente del contexto de la implementación y de la disponibilidad de una institucionalidad adecuada para el levantamiento de la información. La evidencia muestra que un diseño pobre de la estructura de

incentivos puede llevar a que los efectos de estas políticas sean nulos e incluso negativos. Por ello, es relevante un análisis caso a caso que considere a la medición de VA como una de varias mediciones cuantitativas y cualitativas. Además, la investigación sostiene razonablemente la idea de que, si bien los incentivos grupales son más eficientes que los individuales, el tamaño del grupo importa, siendo recomendable que los grupos de profesores definidos para la aplicación de incentivos sean relativamente pequeños, por ejemplo, asignando estos a profesores de un curso o asignatura en vez del colegio completo. Por otro lado, las estructuras de recompensas por metas tienden a ser menos efectivas. En este sentido, son más recomendables los premios a las contribuciones marginales, ya sea a nivel individual o grupal. Por último, para la institucionalidad chilena es de especial importancia evaluar la relación causal de la Evaluación Docente con el resultado de los estudiantes, la dispersión de los puntajes de los docentes sometidos a ésta y la congruencia en el diseño de las pruebas estandarizadas disponibles (SIMCE), si se considera la aplicación de mediciones de VA para evaluar a los docentes del país.

## Referencias

Alvarado, M., G. Cabezas, D. Falck & M. E. Ortega. 2012. "La evaluación docente y sus instrumentos: Discriminación del desempeño docente y asociación con los resultados de los estudiantes". Artículo de investigación, Centro de Estudios Ministerio de Educación de Chile y Programa de Desarrollo de las Naciones Unidas [PNUD], Santiago.

Baker, E. L., Shavelson, R. J., Linn, R. L., Ladd, H. F., Darling-Hammond, L., Shepard, L. A., & Rothstein, R. (2010). Problems with the use of student test scores to evaluate teachers. Washington, DC: *Economic Policy Institute, Briefing Paper* Nº278.

Banerjee, A.V., S. Cole, E. Duflo & L. Linden. 2007. "Remedying Education: Evidence from Two Randomized Experiments in India". *Quarterly Journal of Economics* 122(3): 1235-1264.

Bravo, D., D. Falck, R. González, J. Manzi & C. Peirano. 2008. "La relación entre la evaluación docente y el rendimiento de los alumnos: evidencia para el caso de Chile". Centro de Microdatos, Departamento de Economía, Universidad de Chile.

Carneiro, P., O. Koussihouèdé, N. Lahire, C. Meghir & C. Mommaerts. 2015. "Decentralizing Education Resources: School Grants in Senegal", *NBER Working Paper* Nº21063.

Chetty, R., J. N. Friedman & J. Rockoff. 2014a. "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review*, 104(9): 2593-2632.

Chetty, R., J. N. Friedman & J. Rockoff. 2014b. "Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood." *American Economic Review*, 104(9): 2633-79.

Chetty, R., J. N. Friedman & J. Rockoff. 2015. "Measuring the Impacts of Teachers: Response to Rothstein (2014)", (No. 10768). *CEPR Discussion Papers*.

Chetty, R., J. N. Friedman & J. Rockoff. 2016. "Using Lagged Outcomes to Evaluate Bias in Value-Added Models", (No. w21961). *National Bureau of Economic Research*.

Dee, T. & J. Wyckoff. 2013. "Incentives, Selection, and Teacher Performance: Evidence from IMPACT". *NBER Working Paper* Nº19529.

Glazerman, S. & A. Seifullah. 2012. "An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) after Four Years (Final Report)". Washington, DC: *Mathematica Policy Research*, Inc.

Glewwe, P., N. Ilias & M. Kremer. 2010. "Teacher Incentives". *American Economic Journal: Applied Economics*, American Economic Association, vol. 2(3): 205-27.

Hanushek, E. A. 2011. "The economic value of higher teacher quality". *Economics of Education Review* 30 (3): 466-479.

Hanushek, E. A. & S. G. Rivkin. 2010. "Generalizations about using value-added measures of teacher quality". *The American Economic Review*: 267-271.

Heckman, J. J., R. Pinto & P. A. Savelyev. 2013. "Understanding the mechanisms through which an influential early

childhood program boosted adult outcomes". *American Economic Review*, 103(6): 2052-86.

Imberman, S. A. 2015. "How effective are financial incentives for teachers?". IZA World of Labor.

Imberman, S. A. & M. F. Lovenheim. 2015. "Incentive strength and teacher productivity: Evidence from a group-based teacher incentive pay System". *Review of Economics and Statistics* 97(2): 364-386.

Jacob, Brian A. & Lars Lefgren. 2008. "Can Principals Identify Effective Teachers? Evidence on Subjective Performance Evaluation in Education." *Journal of Labor Economics*, 26(1): 101-36.

Kane, T. J. & D. O. Staiger. 2008. "Estimating teacher impacts on student achievement: An experimental evaluation". *National Bureau of Economic Research, working paper* Nº14607.

Kane, T. J. & D. O. Staiger. 2012. "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains". Research Paper. MET Project. Bill & Melinda Gates Foundation.

Kane, T., D. Staiger & A. Bacher-Hicks. 2014. "Validating Teacher Effect Estimates using Between School Movers: A Replication and Extension of Chetty et al." Harvard University Working Paper.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller & Douglas O. Staiger. 2013. "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." Seattle, WA: Bill & Melinda Gates Foundation.

Lavy, V. 2009. "Performance pay and teachers' effort, productivity, and grading ethics." *American Economic Review* 99(5): 1979–2011.

Martinez, J. F. 2011. "La Evaluación Docente en Chile: Perspectivas de validez". En: Manzi, J., R. Gonzales & Y. Sun. 2011. "La evaluación docente en Chile". MIDE UC.

Manzi, J., R. Gonzales & Y. Sun. 2011. "La evaluación docente en Chile". MIDE UC.

Mihaly, K., et al. 2013. "A composite estimator of effective teaching." Technical report for the Measures of Effective Teaching project. Enero 8.

Muralidharan, K. & V. Sundararaman. 2011. "Teacher performance pay: Experimental evidence from India." *Journal of Political Economy*, 119(1): 39–77.

Murnane, Richard J., John B. Willett & Frank Levy. 1995. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics*, 77(2): 251–66.

Rivkin, Steven G., Eric A. Hanushek & John F. Kain. 2005. "Teachers, Schools, and Academic Achievement." *Econometrica*, 73(2): 417–58.

Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement." *Quarterly Journal of Economics*, 125(1): 175–214.

Rothstein, J. 2014. "Revisiting the impacts of teachers". Trabajo en progreso sin publicar. Disponible en: http://eml.berkeley.edu/~jrothst/workingpapers/rothstein_cfr.pdf.

Springer, M. G., D. Ballou, L. S. Hamilton, V.-N. Le, J. R. Lockwood, D. F. McCaffrey, M. Pepper & B. M. Stecher. 2012. "Final Report: Experimental Evidence from the Project on Incentives in Teaching (POINT)". Nashville, TN: National Center on Performance Incentives.

Todd, P. E. & K. I. Wolpin 2003. "On the specification and estimation of the production function for cognitive achievement". *The Economic Journal* 113(485): F3-F33. **PdR**