No Child Left Behind: retos metodológicos y recomendaciones para la medida del progreso anual adecuado

No Child Left Behind: methodological challenges and recommendations for measuring adequate yearly progress

Yeow Meng Thum
Michigan State University. College of Education. MI, USA

Resumen

Este artículo se ocupa de los retos metodológicos de la rendición de cuentas de estudiantes, escuelas, distritos y estados, tal y como se formula en la nueva ley norteamericana, *No Child Left Behind,* y aporta algunas recomendaciones, basadas en investigaciones recientes, para una aproximación viable de la medida de progreso de las escuelas hacia una meta establecida. Concebido como un documento para la planificación, este artículo busca aportar una plataforma analítica que sea lo suficientemente transparente para que la discusión sobre los procedimientos de medida de la rendición de cuentas que puedan separarse del lado más político del actual debate sobre rendición de cuentas. Se presentan las principales justificaciones para 1) emplear puntuaciones de intervalos, 2) utilizar múltiples resultados, 3) estimar las ganancias en valor añadido a partir de datos longitudinales de los alumnos individuales, 4) exigir modelos basados en la agregación, 5) exigir modelos basados en la inferencia y 6) mantener abierta la *caja negra* de un sistema de rendición de cuentas viable. Dentro del mismo marco teórico, se propone una definición de lo que significa para una escuela «conseguir AYP» en el contexto de la NCLB. Se muestra que esta noción de AYP, denominada «AYP-NCLB», se puede *operacionalizar* como una comparación en cualquier punto temporal de la tasa de crecimiento de una escuela con un mínimo de crecimiento exigido para esa escuela si se quiere

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

que sea «competente» en 2013-2014. El mismo análisis informa de la proporción de los alumnos que son «competentes» en una escuela cada año, que es el principal interés de las aproximaciones de evaluaciones basadas en estándares de referencia.

Palabras clave: sistemas de rendición de cuentas, *No Child Left Behind,* modelos de crecimiento, modelos multinivel multivariados, sistemas de rendición de cuentas.

Abstract

This paper dwells on the methodological challenges for student, school, district, and state accountability as formulated by the new law and provides some recommendations, based on recent research, for a viable approach for measuring progress of schools toward a set target. Conceived as a planning document, this paper aims to provide an analytic platform that will be transparent enough so that the discussion of the procedures for accountability measurement can be better de-coupled from the more contentious policy side of the current school accountability debate. It outlines the principal rationale for (1) employing scale scores, (2) using multiple outcomes, (3) estimating value-added gains from student-level longitudinal performance data, (4) requiring model-based aggregation, (5) requiring model-based inference, and (6) keeping the *black-box* open in a viable accountability system. Within the same framework, it proposes a definition of what it means for a school to 'make AYP' under NCLB. It shows that this notion of AYP, termed 'AYP-NCLB', can be *operationalized* as a comparison at any point in time of a school's growth rate with a minimum growth required of that school if it is expected to be proficient by 2013-14. The same analysis yields the proportion of the students in a school who are 'proficient' each year, the primary interest of standards-referenced approaches to the assessment.

Key Words: accountability system, No Child Left Behind, growth models, multilevel multivariate models.

Introducción

La Ley *No Child Left Behind* de 2001 (NCLB, 2001) representa el papel del gobierno federal mucho más extenso en la educación pública, al caracterizar la nueva ley las provisiones para reforzar la rendición de cuentas del rendimiento académico. Más allá del requerimiento de la medición anual, la ley busca un método para juzgar la eficacia de las escuelas, instaurando una agenda para el progreso último, y estableciendo una

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

secuencia de las consecuencias específicas para el fallo. Con esta legislación, el gobierno federal parece satisfacerse a sí mismo con el papel de un árbitro de los objetivos de rendimiento y progreso, dejando el establecimiento de la base de la evidencia para los juicios, –incluyendo las materias curriculares y la elección de instrumentos de evaluación–, a los estados. El objetivo inmediato de la NCLB parece ser el establecimiento de un conjunto de procedimientos que ayudarán a vincular las evaluaciones a través del tiempo, a través de los sistemas, y con componentes de evaluación externa tales como los de *National Assessment of Educational Progress* (NAEP) para aportar alguna validación del sistema. Un instrumento común de medida para monitorizar el progreso educativo de los niños de la nación puede ser la clave para producir una divisa común en la evaluación de la productividad, su falta es para muchos el principal impedimento para hacer crecer un esfuerzo nacional coherente orientado a la mejora del debate sobre la educación pública.

Este artículo se ocupa de los retos metodológicos de la rendición de cuentas de estudiantes, escuelas, distritos y estados, tal y como se formula en la nueva ley, y aporta algunas recomendaciones, basadas en investigaciones recientes, para una aproximación viable de la medida de progreso de las escuelas hacia una meta establecida. Queda fuera de la perspectiva de este artículo los problemas relacionados con la alineación con el currículum, los estándares y los test o la elección de test alternativos u otras formas de medir. De forma más concreta, me centraré en las siguientes dos cuestiones esenciales de un esquema de medida de la rendición de cuentas utilizable, como son:

- Definir, medir y monitorizar el progreso del rendimiento de valor añadido e informar de la productividad de unidades múltiples y anidadas (subgrupos de alumnos, escuelas, distritos o estados), y
- Definir el progreso anual adecuado (*adequate yearly progress,* AYP) recogido en la NCLB y diseñar un procedimiento para apreciar y para comparar el progreso de las unidades de rendición de cuentas en términos de AYP.

Concebido como un documento para la planificación, este artículo busca aportar una plataforma analítica que sea lo suficientemente transparente para que la discusión sobre los procedimientos de medida de la rendición de cuentas que puedan separarse del lado más político del actual debate sobre rendición de cuentas. Se presentan las principales justificaciones para 1) emplear puntuaciones de intervalos, 2) utilizar múltiples resultados, 3) estimar las ganancias en valor añadido a partir de datos longitudinales de los alumnos individuales, 4) exigir modelos basados en la

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

agregación, 5) exigir modelos basados en la inferencia y 6) mantener abierta la *caja negra* de un sistema de rendición de cuentas viable. Dentro del mismo marco teórico, se propone una definición de lo que significa para una escuela «conseguir AYP» en el contexto de la NCLB. Se muestra que esta noción de AYP, denominada «AYP-NCLB», se puede *operacionalizar* como una comparación en cualquier punto temporal de la tasa de crecimiento de una escuela con un mínimo de crecimiento exigido para esa escuela si se quiere que sea «competente» en 2013-2014. El mismo análisis informa de la proporción de los alumnos que son «competentes» en una escuela cada año, que es el principal interés de las aproximaciones de evaluaciones basadas en estándares de referencia. Además, la estrategia analítica propuesta añade directamente cuestiones referidas a 1) la precisión de las decisiones, 2) la elección del punto de partida para realizar evaluaciones y, 3) las llamadas previsiones de «puerto seguro» en la NCLB.

Como líneas maestras y bosquejos que son, por favor entienda que los argumentos de este artículo están necesariamente abreviados. Las propuestas de este artículo se basan sobretodo en el constante estudio e investigación de muchos académicos pero, en aras de la legibilidad, sólo citaré las fuentes primarias en las que se basa este documento y dejaré al lector consultar las referencias allí contenidas (Thum, 2002; Thum, 2003).

El uso de las puntuaciones de los test

Como en la mayoría de las aplicaciones de la rendición de cuentas, la NCLB cuenta con el uso de evaluaciones estandarizadas de los estudiantes con las que se dirige inmediatamente la atención hacia el consiguiente debate sobre el uso correcto de las puntuaciones de los test. Ha sido largo e inquietante el debate sobre el uso de las puntuaciones procedentes de pruebas estandarizadas para tomar decisiones educativas centradas en las cuestiones sobre la precisión de las puntuaciones de los test, incluso cuando sólo se han usado test válidos y fiables. No obstante, una puntuación de un test puede ser útil si cuidadosamente ponderamos su validez y su precisión. Vean también el reciente *Standards for Educational and Psychological Testing,* 1999.

Mientras que parece claro que ninguna puntuación de un test determinará perfectamente el nivel de rendimiento del estudiante, éste está basado sin embargo en las respuestas de un alumno a una muestra suficientemente amplia de las tareas que definen el dominio. Por tanto, las imágenes que sólo destacan la inherente imprecisión de los resultados

de los test son alarmistas si se pretende que las puntuaciones de los test hagan perfectamente su trabajo. Una puntuación de test, después de todo, es una estimación. Es simplemente una conjetura informada basada, a pesar de todo, en la explícita e imperfecta evidencia del rendimiento. Para cualquier prueba fiable y válida, una medida de cuán (im)precisa puede ser una puntuación se encuentra en su acompañante error típico de medida (*standard error of measurement,* SEM). Sin tener explícitamente en cuenta la imprecisión de las puntuaciones, las inferencias sobre diferencias, tanto si son positivas como negativas, pueden estar sesgadas y las inferencias tenderían a ser demasiado liberales. La cuestión entonces no es si una puntuación de un test puede equivocar la calificación, sino si está sesgada en algún sentido y cuánto, y qué impacto tendrá está imprecisión sobre las decisiones individuales. Por estas razones, recomiendo procedimientos que tengan en cuenta explícitamente el error típico de medida de las puntuaciones.

Definiendo y midiendo la productividad del valor añadido

Un buen entendimiento del cambio en los aprendizajes de un alumno es crítico para mejorar la escolaridad pública. La aproximación del valor añadido al problema de la medida del aprendizaje de los alumnos busca situar el cambio dentro del alumno, aislándolo tan bien como sea posible de los muchos factores omnipresentes relacionados con la historia económica y social del estudiante, y de la agrupación con su escuela y su comunidad. Mientras pueden variar las formulaciones específicas de esta idea básica, y dado que el éxito de las mismas no puede garantizarse en ninguna instancia debido a la relativamente exigente técnica y los requerimientos de calidad de los datos (a la vez que necesarios), una aproximación de valor añadido alberga la mayor promesa para responder a la cuestión:

¿Cómo están aprendiendo los niños en nuestras escuelas?

Sin embargo, la intensidad del actual interés en la medida del valor añadido es relativamente nuevo, los métodos para desarrollar los análisis están relativamente bien establecidos en la literatura sobre métodos de investigación. Thum (2002) revisó recientemente la literatura metodológica sobre la medida del cambio y formuló argumentos para apoyar algunas elecciones críticas en la construcción de un sistema estadístico para observar el progreso académico. Algunas de las principales conclusiones son:

La métrica importa para medir el cambio

En la base de un sistema viable para medir el cambio está una escala de medida de intervalos como métrica deseable para la equiparación. Mientras que el trabajo analítico se viene haciendo en esta escala subyacente, las categorías que reflejan los niveles ordenados de rendimiento pueden usarse para establecer objetivos y para informar de los resultados. La falta de una escala de intervalos en una métrica de equiparación, que es un requisito métrico mínimo para comparaciones válidas del cambio, llevará a imposibles análisis consistentes de rendición de cuentas *–con o sin valor añadido–.* Es la responsabilidad de los productores de test aportar constantemente la necesaria evidencia para sus escalas, haciendo explícito cualquier cambio en los procedimientos, convenciones y asunciones de los modelos que necesariamente son los componentes de la medida estandarizada, así como apoyar el uso adecuado de sus escalas.

Resultados múltiples ayudan

Las medidas múltiples sirven para replicar nuestras lecturas sobre un constructo de rendimiento, no simplemente como un límite o artimaña para intencionalmente presentar una meta confusa. Cuando se despliegan adecuadamente, los sistemas que emplean múltiples medidas nos ayudan a triangular un constructo de rendimiento más general que entendemos está de forma imperfecta representado por una única media. Además, la redundancia de la información en las medidas múltiples también ayuda a reducir el impacto de los errores de medida. Los análisis multivariados, aquellos que tratan todas las puntuaciones de los test como resultados de forma simultánea, aportarán un conjunto más coherente de resultados cuando sean comparados con los intentos racionales de integrar análisis separados de pruebas individuales de las asignaturas.

Para medir el cambio, ganancias estimadas

De las distintas aproximaciones para definir el valor añadido, sólo la ganancia en el nivel del alumno aporta un mapa congruente del cambio en el aprendizaje. La puntuación de la ganancia bruta es simplemente una composición lineal de dos medidas positivamente correlacionadas. Basándonos en el ampliamente aceptado modelo de la «puntuación verdadera», podemos demostrar que, si las medidas compuestas son

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

relativamente precisas, la puntuación de la ganancia bruta tiene una varianza menor que la suma de las varianzas de cada componente de la medida, debido a la correlación entre las puntuaciones verdaderas. La investigación ha demostrado que la fiabilidad de las ganancias dependerá no sólo de la precisión de sus componentes sino también de la distribución de las ganancias de la población, con el resultado de que las puntuaciones de ganancia no son siempre menos precisas que alguno de sus componentes. Por ejemplo, si claramente observamos amplias ganancias que son todas iguales en magnitud, la fiabilidad de las ganancias observadas –una medida normativa de las diferencias en ganancias más allá del ruido del background– es cero.

Y sin embargo, las ganancias brutas pueden no ser inherentemente poco fiables como se pensaba anteriormente, he recomendado que los procedimientos de rendición de cuentas estimen ganancias. Esto se puede conseguir situando todas las puntuaciones de test al mismo nivel como resultados, en lugar de emplear ganancias brutas como punto de partida para el análisis. Es también fácil de demostrar que, dado que este modelo particular de valor añadido emplea al sujeto como su propio control, los factores del nivel individual (como la raza o una beca de ayuda al comedor) que pueden tener un impacto comparable en el rendimiento de los estudiantes en cada medición, ya no predicen las ganancias que se alcanzan. Sin embargo, las ganancias del nivel de aula o de escuela pueden estar correlacionadas con las medidas del aula o de la escuela de esos mismos factores.

Finalmente, la ganancia estimada no tiene ninguna de las dificultades conceptuales y metodológicas que se esperan de las ganancias residuales obtenidas mediante la regresión de las puntuaciones de los estudiantes del postest sobre el pretest. No hace sólo que sus resultados dependan críticamente del agrupamiento particular de la clase o la escuela, si se usara el pretest como control cuando está correlacionado con el resultado, se violarían las asunciones básicas de la regresión lineal (por ejemplo, que los predictores sean fijos y conocidos, y no correlacionados con los residuos).

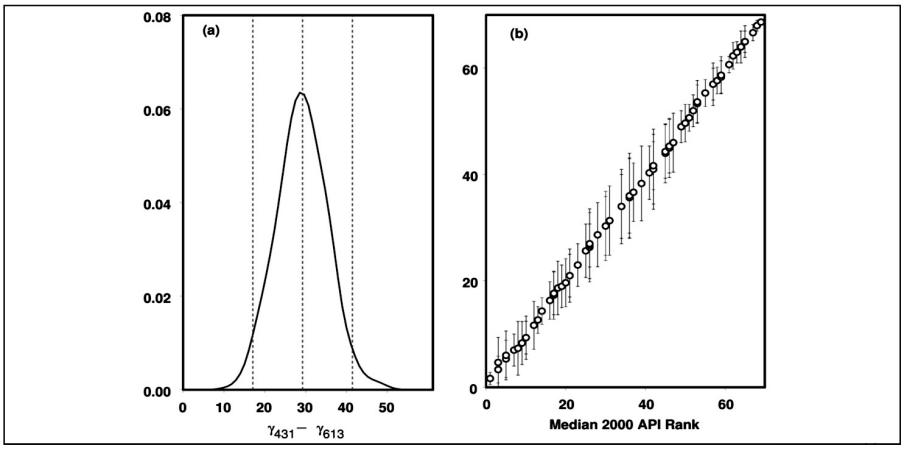
Exigencia de un modelo basado en agregación

Los recientes avances en la investigación sobre la eficacia de las escuelas ha demostrado que la historia sobre el nivel de rendimiento del alumno y de la escuela y el progreso cambia, a veces de forma irreconciliable, cuando promediamos las puntuaciones de los test de varias maneras. Es bien conocido que, por ejemplo, la diferencia entre las medias de las puntuaciones del tercer y cuarto curso no siempre son iguales

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

a la media de las diferencias de las puntuaciones de los alumnos individuales en esos mismos cursos, a menos que el análisis implique a los mismos estudiantes en tercero y cuarto. No sólo la agregación tiene un impacto decisivo en las conclusiones, sino que define la unidad conceptual de qué está siendo medido y, como resultado, esos cambios son los que seguimos. Informar al nivel de la escuela, del distrito o del estado debe realizarse dentro de un marco de modelización coherente a la vez que estadísticamente flexible que comienza con el seguimiento de los cambios en el nivel del estudiante individual y permite una inferencia adecuada de los resultados desagregados.

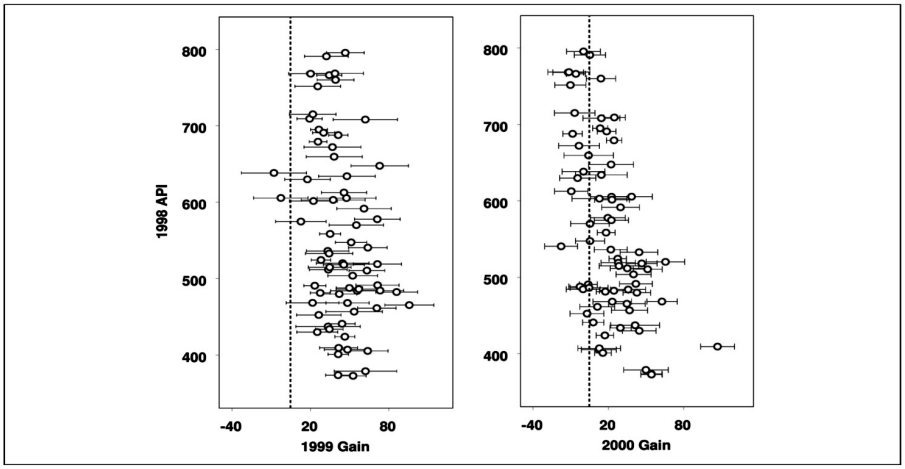
FIGURA I. En California, el API (*Academic Performance Index*) es un índice ponderado compuesto del rendimiento de los estudiantes. a) Comparación de las escuelas 431 y 613 en sus ganancias en API, $\hat{\gamma}_{431}$ y $\hat{\gamma}_{613}$, respectivamente. Las líneas de referencia marcan la diferencia media estimada en el percentil 2,5, en la media y en el percentil 97,5. b) Ordenación (con intervalos de confianza) por la mediana del API en 2000, estableciendo sus intervalos de confianza en el 95%.



Exigencia de un modelo basado en la inferencia

Los sistemas de rendición de cuentas, para ser mínimamente útiles, deben aportar estimaciones defendibles sobre la fiabilidad de sus puntuaciones de productividad. Los cálculos, proyecciones, y ordenaciones del nivel de rendimiento y productividad

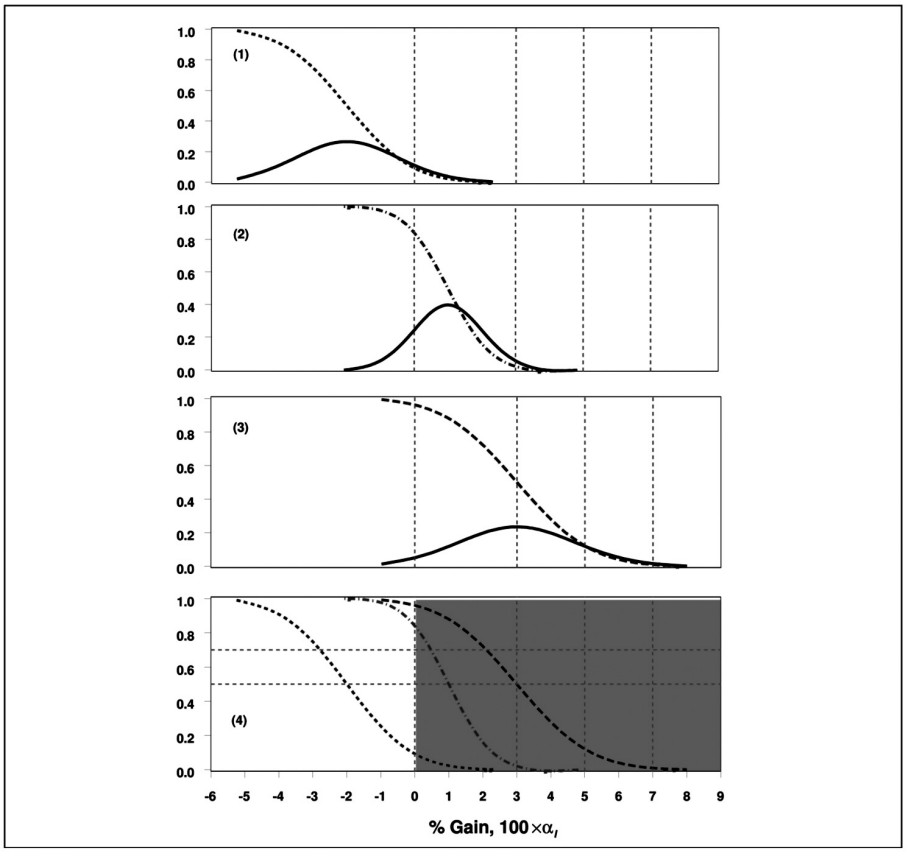
FIGURA II. Estimaciones de las ganancias del API de las escuelas de LBUSD y las estimaciones del intervalo de confianza del 95% en 1999 y 2000 frente a su nivel en API en 1998. Las ganancias en el API de 1999 correlacionan -0,31 (0,12) con el nivel en el API de 1998, mientras que la ganancia del 2000 correlaciona -0,52 (0,09) con el nivel de 1998. Escuelas con el mismo nivel en 1998 ganan menos en 2000 (-0,10 puntos) que en 1999 (-0,06) en promedio. Las líneas verticales de referencia señalan ganancia 0.



no acompañados por un explícito cómputo de las distintas fuentes de medida y variabilidad muestral deben evitarse porque tales incertidumbres impactarían en las decisiones basadas en las estimaciones brutas. Todas las decisiones de alto impacto deben estar cualificadas por afirmaciones inferenciales construidas adecuadamente para representar completamente la dimensión del uso de la información y el grado de precisión. Estar en disposición de ofrecer estimaciones fiables sobre las puntuaciones de productividad es esencial para un sistema de rendición de cuentas defendible. Las figuras I y II son ejemplos, basadas en una cohorte del Distrito Unificado de Long Beach (LBUSD), de un modelo basado en las comparaciones de las ganancias sobre el valor añadido entre dos escuelas y de los ranking de las estimaciones del nivel de las escuelas, respectivamente. Los perfiles de productividad, tales como los que se muestran en las figuras III y IV, representan una aproximación para presentar como un profesor, una escuela o un distrito están progresando y en que nivel estadístico de confianza, *simultáneamente,* dada la evidencia disponible (Thum, 2002).

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

FIGURA III. Los perfiles de productividad para tres escuelas, en los paneles 1, 2 y 3, están superpuestos en el panel 4 para una comparación más fácil. Cada uno está construido a partir de la distribución marginal posterior simulada de la ganancia de la escuela. Un punto en cada línea indica la ganancia estimada 100 x α_i % hecha por una escuela hacia el objetivo de rendimiento (en el eje horizontal) y cómo de fácil es que se obtenga una ganancia observada tan grande como 100 x α_i % en términos de probabilidad (eje vertical).



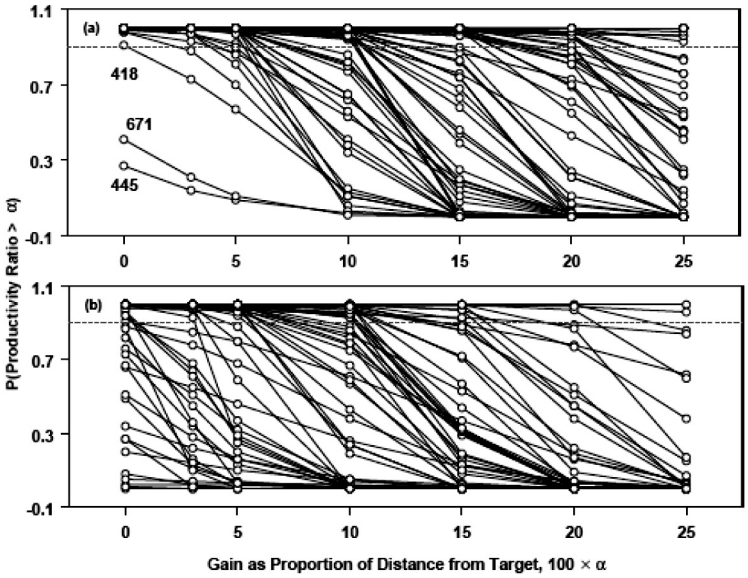
Mantener abierta la «caja negra»

Parafraseando a la profesora Anita A. Summers de Wharton School, nadie necesita abrir la «caja negra» de un sistema potencialmente útil de rendición de cuentas, del mismo modo que ningún conductor necesita entender cómo funciona un coche para estar

cómodo con su uso, *siempre que algún profesional esté encargado de su diseño y seguridad.* No obstante, necesitamos dejar la llave justo encima de la «caja negra» para facilitar un solícito acceso. Nada debe impedir el desarrollo de técnicas convincentes, en tanto que esté lo suficientemente abierta para la revisión entre pares, la evaluación profesional y la auditoria sistemática de su potencial valor o dañaremos el tronco de su desarrollo. Y como la aproximación recibe más y más pruebas realistas en su camino, las inadecuaciones de la metodología o los mitos con respecto a su utilidad práctica, –ambos viejos y nuevos–, serán rápidamente identificados y debidamente superados.

En los párrafos anteriores, he destacado las características de una aproximación profundamente razonada, aproximación que espero sirva para propósitos de diagnóstico y rendición de cuentas, o ambos, mejorando los sistemas que existen actualmente. Estas mismas cualidades la convierten en una fuerte candidata a ser el componente central de un procedimiento de rendición de cuentas. Desgraciadamente, esto también sugiere que la continuada existencia de la mayoría de los sistemas será mucho más difícil de defender desde bases conceptuales y metodológicas.

FIGURA IV. Los perfiles de productividad de LBUSD reflejados en términos de la PSAA ratio de California, para (a) 1999 y (b) 2000. Las escuelas pueden ahora ser más fácilmente comparadas en términos de su productividad para un nivel de precisión determinado y razonable.



Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Progreso anual adecuado

La NCLB exige que todos los estudiantes de las escuelas públicas desde 3º a 8º curso sean competentes en Matemáticas y Comprensión lectora en el año académico 2013-2014, siendo activado el mecanismo de rendición de cuentas en 2005-2006. Esto sugeriría que independientemente del tipo de prueba o del curso específico, el cuerpo de estudiantes estará situado en el nivel de rendimiento de competente en Comprensión lectora y Matemáticas en un tiempo de 12 años. Esto sirve como propósito retórico, pero lo que realmente significa en la práctica es ambiguo. Por ejemplo, no todos los estudiantes cumplirán esta finalidad en el mismo período de tiempo. Y, tal y como la NCLB requiere, si un estudiante comienza la escuela, digamos en 2012-2013, ¿debería ser competente en 2013-2014? ¿Los alumnos que terminaron 8º en 2006 necesitaron ser competentes en ese momento?

Una interpretación razonable podría ser que la NCLB se propone para las escuelas en 2013-2014, no para los alumnos individuales. En el ínterin, son las escuelas las que necesitan mostrar que se han encaminado hacia la competencia en 2013-2014; consecuentemente también necesitan mostrar la importancia de una clara definición de qué significa *lograr una unidad de rendición de cuentas AYP.* La NCLB está primeramente interesada en cómo se mueve cada unidad de rendición de cuentas hacia el objetivo del 100% de competencia, o algún valor que esté aceptablemente próximo a éste, en 2013-2014. La importante cuestión de quién, entre el cuerpo de estudiantes de una escuela, debe ser incluido en la estimación de su productividad en cualquier momento se queda atrás, pero puede ser abordada más tarde.

Muchas de las actuales sugerencias para el AYP giran alrededor de tres ideas (Goertz, 2001). En Texas, las escuelas deben alcanzar un umbral absoluto en rendimiento y en otros criterios. Las metas relativas de crecimiento son empleadas en California. Michigan ofrece un ejemplo en el que el objetivo principal es el descenso de la proporción de estudiantes situados en los niveles más bajos de rendimiento. Todos ellos tienen a su manera un sentido intuitivo y quizá deba ser seguido, privilegiados por algunos estados, al menos en una proporción de sus datos como parte de una estrategia de rendición de cuentas más comprehensiva. Además del intenso interés de muchas agencias estatales, como resultado de la NCLB, está cómo cada una de estas aproximaciones debe ser rearticulada en términos de la NCLB.

No obstante, casi todos los estados establecen una meta de progreso intermedia (mayoritariamente anual) en términos de un nivel de rendimiento o tasa, aunque no todos están claros sobre el marco temporal para lograr este objetivo eventual. Por

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

ejemplo, California requiere que las escuelas ganen un 5% fijo anual con respecto a su distancia entre su API (*Academic Performance Index*) y el objetivo API del estado fijado en 800, sin ningún requisito temporal para alcanzarlo. La NCLB, por el contrario, establece una línea temporal muy clara, 12 años, para que todos alcancen el nivel competente de rendimiento. Para la NCLB, y por tanto el AYP, debe implicar una solución viable para las cuestiones sobre rendición de cuentas esenciales para las escuelas:

> Teniendo en cuenta donde estás en este momento, ¿estás mejorando a un paso que te situará en la meta especificada en el tiempo restante?

Señalaré más adelante cómo una meta explícita y un límite temporal especificado combinan para sugerir una noción del AYP que implica cuestiones de productividad y puntualidad simultáneamente. Específicamente, mi noción de la NCLB sugiere que

> El AYP es definido como la tasa mínima de crecimiento basada en la cantidad de bases que una unidad de rendición de cuentas necesita para completarse y alcanzar la competencia en el tiempo restante. En cualquier punto temporal, las unidades de rendición de cuentas logran el AYP si se mejora en una tasa que iguale o exceda este AYP.

He llamado a este nuevo combinado el «AYP-NCLB» para distinguirlo del otro AYP actualmente en uso. En el AYP-NCLB, evalúo el progreso hacia una meta futura con respecto a una línea base de rendimiento relevante. Como mostraré más adelante, nuestro análisis es fácilmente modificable para aportar una evaluación directa sobre si se puede esperar de una escuela si alcanzará el 100% de competencia en 2013-2014.

Además de su claridad conceptual, las tasas generalmente no «saltan» en la manera en que lo hacen las ganancias anuales. También, y lo único de mi aportación, el sistema de rendición de cuentas está preparado para aportar una respuesta directa a la siguiente cuestión planteada por NCLB:

> Confiamos al P% de que en este momento su escuela está logrando el AYP-NCLB.

Esta afirmación puede ser claramente transmitida desde el perfil de productividad de la escuela, como se muestra en la Figura III.

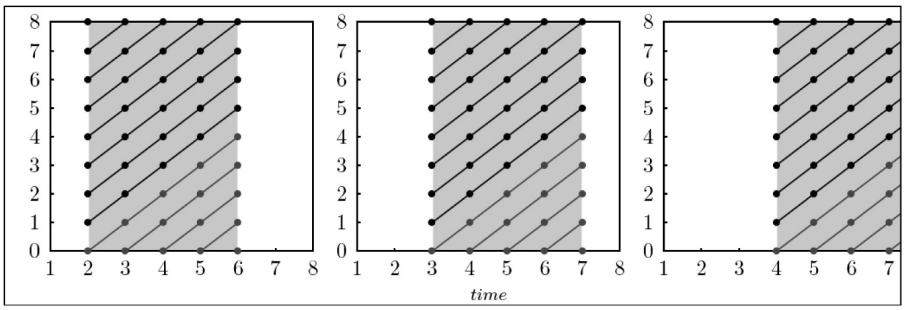
Los datos son una parte de cualquier definición del AYP

Teniendo en cuenta las decisiones con grandes consecuencias implicadas en el seguimiento de la productividad escolar, necesitamos clarificar todavía más la base de evidencia en la que descansan nuestras estimaciones sobre el rendimiento de las escuelas. Primero, se evita sugerir que la productividad de una escuela es un *rasgo* más que un *estado* (en tanto que un rasgo es considerado una cualidad menos efímera que un estado), necesitamos que el rendimiento de la escuela no esté sólo afectado por los muchos factores relacionados con su composición y recursos, siendo especialmente crítico también reconocer que la productividad de una escuela está circunscrita en el tiempo.

Además, cuando se describe la productividad de una escuela en 2006, por ejemplo, necesitamos un mínimo de claridad tal que nuestras evaluaciones estén basadas en una evidencia disponible entre 2002 y 2006. Otra regla puede ser que se emplee un diseño de bloques móviles (*rolling block*), como el mostrado en la Figura V, si creemos que los datos antiguos pueden no ser relevantes dadas las actuales condiciones de la escuela. Creo que esta práctica es especialmente tentadora porque esperamos que los modelos analíticos y los diseños de datos varíen en los distintos sistemas de rendición de cuentas. De forma específica, sugiero que el bloque de 2002-2006 de los datos de evaluación de los estudiantes de una escuela sean analizados simultáneamente con modelos multivariados mixtos que siga simultáneamente todas las cohortes en cada bloque de datos, cada uno de los cuales es representado por la traza desde abajo a la izquierda hasta arriba a la derecha en la Figura V. Después de determinar que nuestro modelo reproduce adecuadamente los datos, podemos entonces calcular los valores ajustados con las estimaciones de su precisión para todos los nodos, cada materia, curso y año (y subgrupo) específico representado tendrá su estatus, ganancia y tasa de crecimiento. Estos resultados aportan los estadísticos necesarios para personalizar las comparaciones que responden a cuestiones de rendición de cuentas distintas. Por ejemplo, podemos seguir el nivel de rendimiento anual realizado por el tercer curso de una escuela a través del tiempo para tener un sentido del progreso del tercer curso en la escuela. Los cambios en la productividad para el tercer curso en la escuela pueden evaluarse comparando las tasas de crecimiento para distintas cohortes longitudinales cada año. Además, el procedimiento también se acomodará a las comparaciones del crecimiento de la escuela o productividad que tienen en cuenta diversas características de la composición de los estudiantes y de la escuela. Los detalles del modelo para utilizar esta estructura de datos en un sistema de rendición de cuentas como la NCLB se presentan a continuación. Desde mi punto de vista, esta aproximación

es también fácilmente adaptable para sistemas que prefieran el seguimiento de índices de puntuaciones, como el californiano API.

FIGURA V. La productividad de una escuela en cada punto temporal debe basarse en un diseño explícito de base de datos de evaluaciones (por ejemplo, 5 años). En eje vertical, el grado 0 comienza en la Escuela Infantil. Nótese que cada traza representa una cohorte de estudiantes longitudinal diferente, los vectores reales de las puntuaciones de los test de los alumnos contendrán valores perdidos.



Análisis frente a unidades de rendición de cuentas

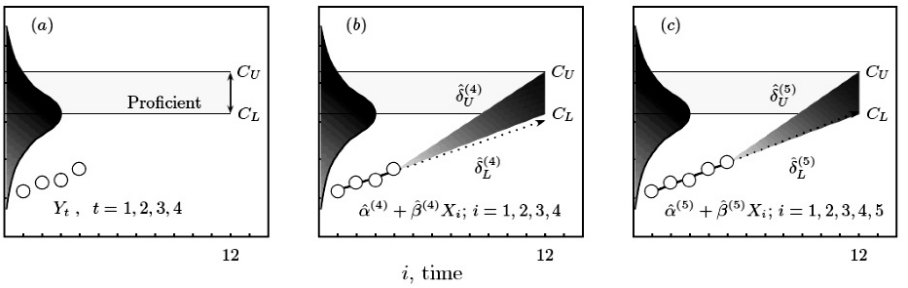
Y, finalmente, necesitamos añadir algunas preocupaciones serias aunque por suerte erróneas desde mi punto de vista, expresadas por muchos de los patrocinadores de los datos longitudinales para la rendición de cuentas. Mientras mi lectura de la NCLB identifica a la escuela y sus subunidades como elementos últimos de la rendición de cuentas, *esto no implica necesariamente que sólo los resultados del nivel de la escuela sean relevantes.* Mi procedimiento propuesto, como señalaba, empleará una base de datos longitudinal de alumnos para caracterizar el AYP-NCLB, en las que *la unidad de análisis sigue siendo (apropiadamente) el estudiante individual, y las unidades de rendición de cuentas son subgrupos de estudiantes y la escuela.*

Logrando el AYP bajo la NCLB

Para medir la productividad en términos del AYP-NCLB, he descrito ampliamente una aproximación para medir el progreso hacia una meta recientemente en Thum (2002) y Thum (2003). Aquí, sólo aportaré un esquema, suprimiendo la notación que sería

necesaria para representar los múltiples resultados y las múltiples unidades de análisis completamente. Para el propósito limitado de este artículo, la lógica esencial está bosquejada con la ayuda de los paneles (a), (b) y (c) de la Figura VI, sólo para una escuela en una única prueba y para un único curso. Se presenta a continuación el tratamiento completo para múltiples medidas y múltiples criterios, incluyendo la sutileza relatada de la forma de la función de predicción y del heterogéneo error de varianza dentro de un modelo para datos anidados.

FIGURA VI. En cualquier punto temporal, *t*, la productividad de una escuela es medida simultáneamente junto con su AYP. Los logros en escala de intervalos están dibujados en el eje vertical frente al tiempo en el eje horizontal (ver el texto para mayores explicaciones).



Puntuaciones y metas

La distribución de puntuaciones para la población se muestra en el fondo de cada gráfica de la figura VI. Las líneas de referencia horizontal marcan las puntuaciones de corte superior e inferior para la meta de rendimiento de «competente», denotados por C_L y C_U respectivamente. Los análisis se realizarán utilizando puntuaciones de intervalos, incluso cuando *algunas lecturas erróneas de la legislación puedan sugerir que los análisis implicados necesiten comenzar y terminar con categorías de rendimiento.* Ya que utilizamos estándares de rendimiento que definen las puntuaciones de logro original en intervalos, aportamos una evaluación más consistente internamente de lo que significa, por ejemplo, ser «competente» y, como resultado, se destaca la relevancia directa de cómo se han establecido los estándares de rendimiento. Se recomienda también una alternativa atractiva más consistente internamente cuando se comparan los estándares de rendimiento que están impuestos internamente, tal y como fija un crecimiento anual fijo del 5% en el API de California, que no están vinculados tan próximamente a los dominios que están siendo evaluados.

Rendimiento y productividad

En el panel (a) de la Figura VI, llamamos «o» al progreso escolar a un resultado de una escuela (Y_1, Y_2, Y_3, Y_4) para los primeros 4 años. El límite temporal, de acuerdo con la NCLB, es $t = 12$. Utilizando un modelo lineal simple, por ejemplo, podremos aproximar cómo está rindiendo la escuela en el momento t, estimando $\underline{\hat{Y}_t}$ a partir de

$$\hat{\alpha}^{(t)} + \hat{\beta}^{(t)} X_i \ .$$

Si el predictor X_i cifrado en el tiempo de tal manera que $X_i = i - t$ entonces $\hat{\alpha}^{(t)}$ da una estimación directa de $\underline{\hat{Y}_t}$. Nótese que la tasa de crecimiento, $\hat{\beta}^{(t)}$, es simplemente una ganancia media estimada, que en tanto que medida de rendimiento, es menos susceptible a fuertes fluctuaciones comúnmente observadas en las ganancias de año en año. Cuando examinamos la conducta de $\hat{\beta}^{(t)}$ a través del tiempo, estamos estudiando la productividad de la escuela para un marco relevante de tiempo.

Definición del AYP-NCLB

Considerando el tiempo restante 12-*t*, la escuela necesitará crecer en una tasa igual a

$$\hat{\delta}_L^{(t)} = \frac{C_L - \hat{Y}_t}{12 - t}$$

basada en nuestra mejor lectura de dónde está la escuela en el tiempo t para alcanzar la meta, limitada por abajo por el punto de corte C_L en $t = 12$. De forma similar,

$$\hat{\delta}_U^{(t)} = \frac{C_U - \hat{Y}_t}{12 - t}$$

aporta la tasa necesaria para superar la competencia. $\hat{\delta}_L^{(t)}$ y $\hat{\delta}_U^{(t)}$ conforman los límites inferior y superior respectivamente de la «raya» de nuestra mejor estimación del nivel actual de rendimiento de la escuela con respecto a la meta temporal situada en $t = 12$. El Panel (b) y el Panel (c) describen las tasas de crecimiento de la escuela y el AYP-NCLB en $t = 4$ y $t = 5$, respectivamente. Una implicación inmediata para casi todas las formulaciones existentes del AYP es que, dado que la meta última es fija en la NCLB, la meta intermedia que es

El AYP-NCLB cambia con el tiempo.

En tanto que seamos capaces de determinar mejor y mejor el progreso de la escuela a través del tiempo, la práctica común de establecer *una tasa anual fija para un período de tiempo extenso tendrá menos sentido.* En definitiva, para ser específico en cada punto temporal, mi razonamiento más bien sugiere que el AYP-NCLB es

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

una unidad específica de rendición de cuentas y también es específica para un test. La agregación de test y de condiciones subagrupadas aportará la estimación apropiada para evaluar cómo varias unidades de rendición de cuentas satisfacen la NCLB.

Alcanzando el AYP-NCLB

A pesar de que podamos calcular en cualquier momento la proporción de estudiantes que alcanzan la competencia, siento que esta imagen sea poco representativa de la tendencia de crecimiento escolar y sugiero en su lugar la siguiente alternativa. Dada nuestra mejor estimación de productividad en el momento t, por ejemplo $\hat{\beta}^{(t)}$, la escuela alcanzará el AYP-NCLB si

$$\hat{\beta}^{(t)} \geq \hat{\delta}_L^{(t)} \ .$$

Entonces, si una escuela alcanza el AYP-NCLB implica una comparación de tasas de crecimiento. Esta comparación aporta una estimación del eventual rendimiento de la escuela. Esto es, esperamos que la escuela rinda en el nivel de competente en 2013-2014 si

$$\hat{\delta}_U^{(t)} \geq \hat{\beta}^{(t)} \geq \hat{\delta}_L^{(t)} \ .$$

En el Panel (b) y el Panel (c), las proyecciones para alcanzar eventualmente la meta están dadas por las líneas negras con puntos. En nuestro ejemplo, en $t = 4$, la escuela parece no alcanzar el AYP-NCLB. En $t = 5$, la escuela parece alcanzar el AYP-NCLB. Tener a la escuela retrasada otra vez en $t = 5$, podría ser objeto de alguna forma de intervención. Parece claro que, en el AYP-NCLB, continuamente evaluamos el progreso hacia una meta futura de rendimiento fija con respecto a una línea base de rendimiento relevante. Debe estar también claro que comparaciones en múltiples metas, como por ejemplo las definidas para otros niveles de rendimiento o para subgrupos específicos de niveles de rendimiento, son también fácilmente desarrollables.

Confianza en la decisión

Nuestros resultados lejos de confiar en la comparación de tasas de crecimiento estimada, representan, por un lado, cómo está rindiendo la escuela, $\hat{\beta}^{(t)}$, y por otro sirve como punto de referencia interino del AYP-NCLB, $\hat{\delta}_L^{(t)}$. Dado que ambos son estimaciones, es importante, como he señalado detalladamente, para caracterizar el nivel de certeza vinculado a sus comparaciones después de tener en cuenta todas las fuentes de variación conocidas. Esto es especialmente claro en nuestro ejemplo para $t = 5$, donde los resultados parecen –para algunos decisores al menos– demasiado próximos para llamar la atención.

De forma más específica, desearíamos conocer cuan plausible es que una escuela logre el AYP-NCLB en t, o

$$\text{Prob}\left(\hat{\beta}^{(t)} \geq \hat{\delta}_L^{(t)} \mid \mathbf{Y}\right).$$

De forma similar,

$$\text{Prob}\left(\hat{\delta}_U^{(t)} \geq \hat{\beta}^{(t)} \geq \hat{\delta}_L^{(t)} \mid \mathbf{Y}\right)$$

aporta la probabilidad de que la escuela sea competente cuando llegue la fecha límite en $t = 12$. Al acordarse que las altas demandas están vinculadas a nuestras decisiones de rendición de cuentas y dado un razonable consenso sobre el grado de seguridad que tenemos antes de realizar afirmaciones sobre si una escuela alcanza o no el AYP-NCLB, podemos seleccionar un rango de niveles de confianza, por ejemplo 70%, 80% o 90% para ayudarnos a llegar a una decisión basada en los datos. Si deseáramos tener mayor certeza sobre si una escuela lograra el criterio AYP-NCLB en el momento t, podríamos seleccionar un nivel de confianza del 90%, entonces la escuela alcanzará el AYP-NCLB si

$$\text{Prob}\left(\hat{\beta}^{(t)} \geq \hat{\delta}_L^{(t)} \mid \mathbf{Y}\right) \geq 0.90 \ .$$

Si nos sentimos todavía incómodos en la selección de un nivel de confianza particular, podríamos emplear el perfil de productividad descrito previamente en la Figura III para ayudarnos a conducir una afirmación más precisa sobre la confianza estadística de nuestra decisión para un abanico deseable de niveles de confianza.

Evaluando los programas de reconocimiento

Las decisiones de rendición de cuentas son difíciles, y por tanto, los análisis cuidadosos de datos son críticos, precisamente porque el verdadero nivel de mejora es desconocido. Sin embargo, cuando las estimaciones de productividad escolares están avaladas por estimaciones fiables en la manera descrita previamente, apoyan firmes acciones de rendición de cuentas (premios o sanciones) en el nivel de escuelas. Por ejemplo, si se otorga un premio a una escuela, tendremos al menos el 90% de certidumbre de que la escuela logra o excede su meta de crecimiento, habiendo aportado una base estadística sólida para cómo son reconocidas las escuelas. De forma similar, se puede realizar para subgrupos específicos, o para cualquier combinación arbitraria de subgrupos. Es interesante destacar que este procedimiento también puede servir

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

como *estándar de oro,* en tanto que no haya otro disponible, de los varios regímenes de reconocimiento. Como ejemplo obvio pero cautivador, podemos entonces evaluar la precisión del examen de un programa de premios alternativo que no tenga en cuenta la precisión y la estimación de la productividad de una escuela si está basado en la comparación del rendimiento escolar con una meta explícita, usando bien una media o una estimación de un porcentaje de corte (PAC). Por último, sugiero, sin desarrollarlo, que esta aproximación para establecer y estimar la precisión de una decisión de rendición de cuentas rodea la especulación directa con respecto al número mágico del tamaño del grupo, el llamado «tamaño mínimo del grupo», exigido para tomar decisiones «estadísticamente fiables».

Puntos de inicio alternativos

Se han propuesto varias alternativas en la legislación para su uso como líneas de referencia. Por ejemplo, utilizar los datos de 2001-2002 puede considerarse como emplear el nivel predicho de rendimiento en el momento *t* del subgrupo con más bajo rendimiento dentro de una escuela, o la escuela con rendimiento más bajo de un distrito, o el nivel medio predicho de un subgrupo de escuelas en el sistema en lugar de $\underline{\hat{Y}_t}$ para la definición de $\hat{\delta}_L^{(t)}$. De hecho, la comparación de cada escuela con líneas de base múltiples, algunas de las cuales puede ser importantes para algunos subgrupos de alumnos, no plantea cargas adicionales para nuestros análisis.

El APY-NCLB y el % de competentes

Aún cuando he designado en el AYP-NCLB una herramienta de pronóstico en términos del rendimiento medio esperado de los alumnos de una escuela en un momento temporal, se puede fácilmente *informar* de los análisis en términos del nivel de rendimiento estimado para un alumno individual. Supongamos que $\underline{\hat{Y}_t}$ denomina a la puntuación predicha para el estudiante *i* de una escuela en el momento *t*. Entonces el estudiante i tiene una probabilidad

$$\text{Prob}\left(\hat{Y}_{it} \geq C_L \mid \mathbf{Y}\right)$$

de ser al menos «competente» y el porcentaje estimado de alumnos en una escuela en el momento *t* que son al menos «competentes» es simplemente

Revista de Educación, 348. Enero-abril 2009, pp. 67-90
Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

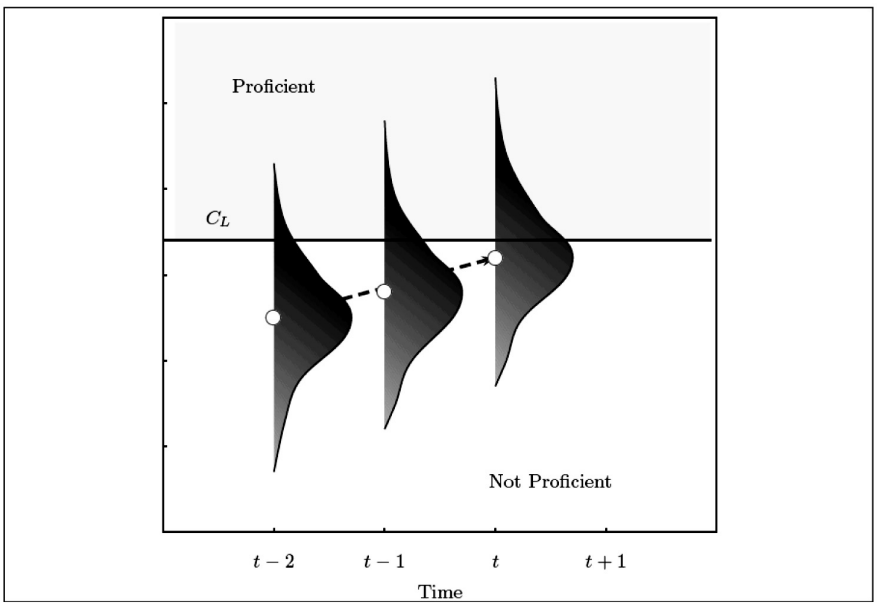
$$\hat{P}_{jt} = \frac{100}{n_t} \sum_{i=1}^{n_t} \text{Prob}\left(\hat{Y}_{it} \geq C_L \mid \mathbf{Y}\right),$$

donde n_t es el número de estudiantes en la escuela en el momento t (véase la Figura VII). Al utilizar la distribución del rendimiento de la escuela en dos años consecutivos, t y t-1, podemos añadir directamente la llamada provisión de «puerto seguro» estimando

$$\text{Prob}\left(\hat{P}_{jt} - \hat{P}_{j,t-1} \geq 10\% \mid \mathbf{Y}\right),$$

para evaluar qué verosimilitud puede tener que haya un decremento de al menos un 10% de alumnos no competentes en la escuela en un periodo de dos años. No hace falta decir que las conclusiones basadas en el AYP-NCLB y los diversos criterios del porcentaje de alumnos pueden divergir en algunos casos. Creo que el último es más conservador y duro para estimar con niveles razonables de precisión.

FIGURA VII. Dadas nuestras estimaciones de crecimiento escolar en cualquier punto temporal, se puede seguir simultáneamente la reducción en el porcentaje de estudiantes no competentes de una escuela. Para cada año, la proporción de alumnos no competentes es la proporción de la distribución por debajo de C_L.



Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Puesto que la mayoría del lenguaje empleado en esta legislación para describir el progreso está en términos de esta distribución, muchas agencias sienten comprensiblemente la responsabilidad de ofrecer este resumen como punto inicial y final de su sistema de rendición de cuentas. Ya sabemos que cuando empleamos categorías de rendimiento que se obtienen por la politomización de las escalas originales de puntuaciones, estamos despreciando información importante, informando de menores, aunque significativos, cambios en los aprendizajes que permanecerán indetectables. Como es también evidente ahora, los análisis adecuados pueden comenzar con puntuaciones de intervalos longitudinales de los estudiantes y los aspectos de los resultados pueden presentarse en términos de la distribución del rendimiento de la escuela. Comenzando con las distribuciones transversales observadas del rendimiento de la escuela tendríamos extremadamente mal representado el carácter longitudinal del crecimiento y del cambio en los aprendizajes de los alumnos.

Resumen y conclusiones

Tanto si está establecido explícitamente o no, es importante reconocer que cada análisis implica un modelo y debemos escudriñar con cuidado cualquier análisis que parezca libre de uno. En este artículo, he aportado un boceto de un procedimiento de rendición de cuentas que responde a la NCLB. El sistema propuesto necesita para operar correctamente comenzar con datos longitudinales del logro de los alumnos. Esto requiere que una prueba basada en estándares con logro expresado en una escala de intervalos no sea una restricción real, en ausencia de alternativas reales. Los que desarrollan las pruebas, que deben jugar un papel más activo para apoyar el renovado esfuerzo sobre la rendición de cuentas en toda la nación, deben aportar garantías periódicas de que las puntuaciones de sus test están equiparadas por cursos y a través del tiempo. Deben también ayudar a clarificar los estándares de rendimiento para sus pruebas, por ejemplo, ¿cuáles son las correlaciones del rendimiento con el nivel de «competencia» en sus test? Debe dirigirse un mayor asesoramiento hacia la tutela del uso adecuado de las puntuaciones de sus test. Con test juiciosos medidos en una escala deseable, los elementos de los testados modelos multinivel multivariados pueden formar la esencia analítica para la estimación del crecimiento del rendimiento, para lo que he superpuesto una nueva formulación de procedimientos para hacer inferencias razonables sobre si una escuela alcanza o no el AYP. Concluyo brevemente añadiendo tres preocupaciones ampliamente expresadas:

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

- Mientras la NCLB configure sus objetivos de logro en términos de categorías de rendimiento, mantengo que esto simplemente facilitará la comunicación siempre que no altere nuestro foco analítico de los resultados continuos de los estudiantes. Por las razones previamente señaladas, tengo serias dudas acerca de las bases conceptuales de las recientes recomendaciones llamadas *evaluaciones referidas a estándares* frente a aproximaciones más «tradicionales» (por ejemplo en Schwarz et al., 2001). Por qué trabajar con información rebajada a categorías de rendimiento cuando tenemos múltiples lecturas del rendimiento de los estudiantes que están sintetizadas por una medida continua de intervalos. No se debe confundir los principios básicos de análisis con la simple conveniencia del informe para construir un sistema de rendición de cuentas sólido. El rendimiento en un criterio y el rendimiento relativo a un conjunto de normas son informaciones complementarias que son igualmente importantes para la salud de la educación pública.
- Necesitamos también evitar agregaciones simplistas de estudiantes inferidas al nivel de escuela porque, sin un modelo explícito de la naturaleza anidada de los datos, las informaciones importantes referidas a los test y las fuentes de variación muestrales se perderán sin posibilidad de recuperación. Y como ya he señalado antes, esto es bastante contrario a muchas de las actuales lecturas de la NCLB: sólo porque la NCLB establece explícitamente metas para las escuelas esto no significa que los resultados de las escuelas deban comenzar por estos análisis. Recomiendo estimaciones del nivel de escuelas sintetizadas a partir de datos longitudinales de los estudiantes y hacer entonces inferencias en función de estos análisis.
- Los estándares altos están bien y son buenos, pero los análisis preliminares en todo el país cuestionan si los objetivos de la NCLB son alcanzables en la práctica. Y esto no quiere decir que debemos establecer objetivos que juzguemos como alcanzables, ya que se puede fácilmente cuestionar su credibilidad. Nótese que, como un estándar que es, el AYP-NCLB establece sólo una senda para buscar un estándar de rendimiento en una determinada cantidad de tiempo. Y mientras podamos tener el impulso para altos estándares por su valor motivacional, y aún cuando el AYP-NCLB aporta un estándar útil para el progreso desde nuestro punto de vista, ciertamente es más necesario conocer el nivel de ganancias de aprendizaje que es posible en muchos contextos de evaluación para llegar a estándares razonables de crecimiento. En este momento, coincido con que las ganancias normativas, como nivel sucesivo de norma, son obviamente piezas que aún están ausentes en el puzzle de la rendición de cuentas.

Fecha de entrada: 29-07-08 Fecha de aceptación: 24-09-08

Referencias bibliográficas

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Goertz, M. E. & Duffy, M., with Le Floch (2001). *Assessment and Accountability Systems in the 50 States: 1999-2000.* A CPRE Report.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Schwarz, R. D., Yen, W. M., & Schafer, W. D. (2001). The challenge and attainability of goals for adequate yearly progress. *Educational Measurement: Issues and Practice,* 20, 26–33.

Thum, Y. M. (2002). Measuring Student and School Progress with the California API. *CSE Technical Report 578.* Los Angeles: Center for Research on Evaluation, Standards, and Student Testing, UCLA.

Thum, Y. M. (2003). Measuring Progress towards a Goal: Estimating Teacher Productivity using a Multivariate Multilevel Model for Value-Added Analysis. *Sociological Methods & Research,* 32, 153-207.

Dirección de contacto: Yeow Meng Thum. Measurement and Quantitative Methods, Counseling Educational Psychology and Special Education, Erickson Hall, Michigan State University, College of Education, East Lansing, 48824 Michigan. E-mail: thum@msu.edu