Programa de Promoción de la Reforma Educativa en América Latina y el Caribe

Grupo de Trabajo sobre Estándares y Evaluación

Responsabilización basada en estándares. Diez sugerencias

Robert L. Linn

Grupo de Análisis para el Desarrollo

RESPONSABILIZACIÓN BASADA EN ESTÁNDARES. DIEZ SUGERENCIAS.

Robert L. Linn[1]

El presente artículo está basado en el CSE Technical Report 490: Assessments and Accountability, de 1998. Ha sido traducido con autorización de sus editores: CRESST.

La historia ha mostrado que la medición es un instrumento de amplio uso para la responsabilización y reforma por un número de razones clave, incluyendo:

1. **Las pruebas son relativamente económicas**

En comparación con cambios que involucran aumento del tiempo de instrucción, reducción del tamaño de las clases, formación y atracción de mejores profesores, la evaluación tiene costos bajos.

2. **Los cambios en la medición pueden implementarse relativamente rápido**

Otras reformas escolares pueden tomar muchos años para implementarse, y puede tomar más aun el saber si han mejorado la escuela.

3. **Los resultados de las pruebas son visibles y llaman la atención de los medios**

Resultados pobres en el primer año de un nuevo programa de medición suelen verse seguidos por el incremento de los puntajes en los años siguientes, dando la apariencia de que las escuelas están mejorando.

4. **La medición puede generar otros cambios que serían difíciles de legislar**

La investigación ha demostrado que los requerimientos de medición y evaluación a nivel estatal o distrital han motivado cambios en el currículo y en la enseñanza en los niveles de la escuela y el aula. Es mucho más difícil legislar directamente cambios en el aula.

Desafortunadamente, cuando las pruebas se usan para tomar decisiones importantes sobre las escuelas y los estudiantes, estos rasgos atractivos dan lugar a problemas inesperados. Los resultados de las pruebas pueden ser incompletos o engañosos, llevando a malas decisiones de política. Sin embargo, la necesidad que tiene el proceso de formulación de políticas de contar con información rápida sobre el progreso de los estudiantes y la calidad de las escuelas asegura un continuado alto interés en la medición educacional.

---

[1] Robert L. Linn es co-director del Centro Nacional para la Investigación sobre Evaluación, Estándares y Medición del Logro Estudiantil y Profesor Distinguido de Educación en la Universidad de Colorado en Boulder. Es el actual Presidente del Consejo Nacional de Investigación en Medición y Evaluación.

Sistemas de evaluación basados en estándares

Un rasgo clave de los esfuerzos de la actual reforma escolar es la creación de estándares educacionales, con el gobierno federal alentando a los estados a desarrollar estándares de contenido y desempeño desafiantes. Los sistemas de evaluación basados en estándares se han convertido rápidamente en el elemento central de muchos programas de reforma estatales, en estados tales como Kentucky o Maryland. Otros estados como Colorado o Missouri están en plena implementación de sus propias evaluaciones basadas en estándares. Ya se ha encontrado que estos sistemas se enfrentan a los mismos retos que anteriores programas de evaluación y a algunos nuevos. Por ejemplo:

1. Los estándares educacionales a nivel nacional, estatal y distrital son a menudo inconsistentes

Revisiones de los estándares de contenido estatales (Education Week, 1997; Lerner, 1998, Olson, 1998, Raimi & Braden, 1998) revelan que varían en un rango de muy fuertes a muy débiles. Diferentes jueces a menudo dan diferentes valoraciones a los mismos estándares, lo que añade al problema.

2. La manera en que se formulan y miden los estándares importa

Tanto la elección de "qué" es medido como la calidad de los estándares y las evaluaciones son importantes. La tabla 1 reporta importantes diferencias en el rendimiento de los estudiantes en las áreas de geografía, historia, matemáticas y lectura, medidas por la Evaluación Nacional de Progreso Educacional (NAEP), "la libreta de calificaciones de la nación".
En la tabla 1, ¿por qué solo el 9% de las estudiantes mujeres están alcanzando un nivel de proficiencia en historia, mientras que el 43% está alcanzando ese mismo nivel en lectura? Si bien las diferencias podrían ser efectivamente diferencias en rendimiento, es mucho más probable que se deban a cómo se formularon los estándares o a la precisión de las evaluaciones al medir sus respectivas áreas.

Una evaluación solo en geografía mostraría que más hombres (32%) que mujeres (22%) alcanzan el nivel de proficiencia mientras que lo contrario sería verdad para una evaluación solo en lectura, con 29% de los hombres proficientes y 43% de las mujeres en ese nivel. Además, la elección de diferentes combinaciones de las 4 pruebas podría producir resultados que fueran casi equivalentes para hombres y mujeres o resultados que favorecerían a un grupo sobre otro. La elección de qué se va a medir puede también alterar las diferencias aparentes en el rendimiento de grupos étnico - raciales o de grupos formados sobre la base de otras características.

Tabla 1

Diferencias en rendimiento según Materia y Género en el NAEP*

Materia	Hombres	Mujeres	Diferencia (H-M)
Geografía (1994)	32	22	10
Historia (1994)	12	9	3
Matemáticas	18	14	4
(1996)	29	43	-14
Lectura (1994)			

* Porcentaje de estudiantes en o sobre el nivel de proficiencia de las pruebas del National Assessment of Governing Board para 12avo grado.

3. A quiénes se incluye o excluye de las evaluaciones puede producir resultados diferentes

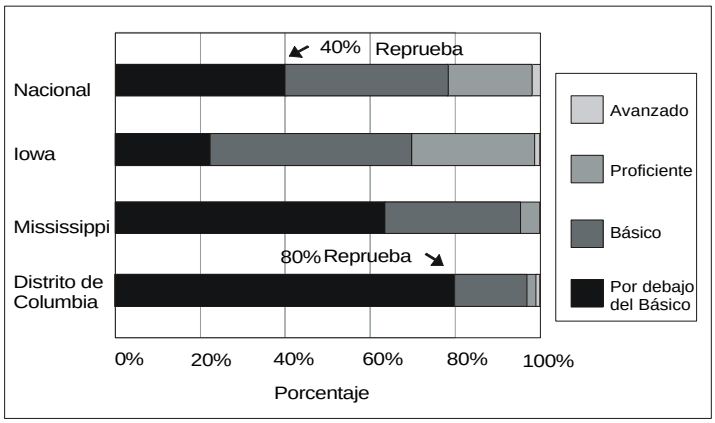
Debido a los requerimientos del programa Título I[2], la reforma basada en estándares enfatiza la inclusión en los programas de evaluación a gran escala, de estudiantes con necesidades especiales y estudiantes cuya lengua materna no es el inglés. La medición brinda información importante a quienes toman decisiones, a los docentes en todos los niveles y a los padres, sobre cómo les está yendo a los alumnos. Sin embargo, la inclusión puede llevarse a extremos. Por ejemplo, evaluar a los alumnos en un idioma que no comprenden producirá puntajes bajos inexactos en las pruebas. Por otro lado, excluir a demasiados estudiantes producirá puntajes inflados. Los retos de la inclusión de todos los estudiantes son difíciles, pero esenciales para un sistema de evaluación creíble.

4. Exigir a todos los estudiantes el logro de los mismos estándares altos dará lugar a tasas altas de retención y fracaso que resultarán inaceptables

El gráfico 1 muestra que cerca del 40% de los estudiantes americanos no alcanzó el nivel básico en la prueba de matemáticas de 1996 para octavo grado del NAEP. ¿Estamos preparados como nación para retener [ese grado] hasta 40% de nuestros estudiantes a nivel nacional, u 80% en algunos distritos? Hacerlo llevaría a grandes cuestionamientos políticos y legales.

[2] Título I (*Title I)* es un amplio programa del gobierno federal de los EEUU de apoyo a iniciativas de sus estados y localidades, de naturaleza compensatoria para grupos de escolares en riesgo, que incluye programas de instrucción remedial, de mejoramiento de la enseñanza, de introducción de innovaciones y otros – NT.

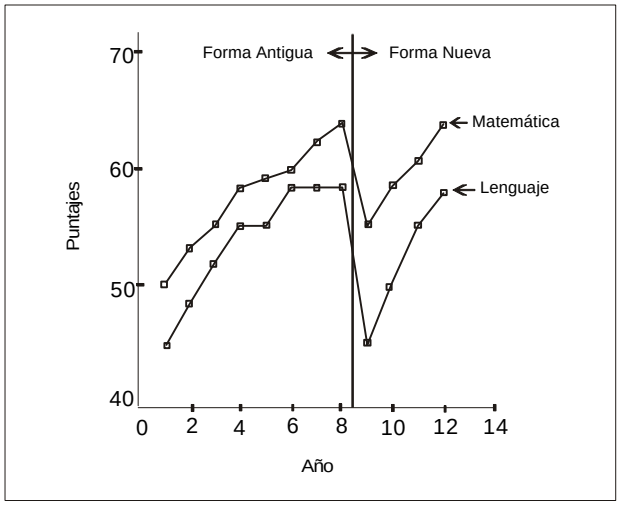
Gráfico 1
NAEP 1996 - 8vo grado: Niveles de logros en matemáticas[3]



5. Los incrementos en los puntajes no reflejan necesariamente verdaderas mejoras

La investigación ha mostrado continuamente que incrementos en los puntajes de las pruebas recién implementadas reflejan otros factores y no [sólo] el incremento en el logro estudiantil. Los incrementos son frecuentemente resultado de que los docentes "enseñan para la nueva prueba" o de que se continúan usando normas de pruebas antiguas (ver gráfico 2). Las evaluaciones basadas en estándares no tienen una mejor capacidad que otros formatos de pruebas para corregir este problema.

Gráfico 2
Resultados de cambiar a una nueva forma de pruebas[4]



Nótese que después de un período de incremento en los puntajes de las pruebas, se introduce una nueva forma de pruebas entre los años 8 y 9.

[3] Basado en Reese et al., 1997.
[4] Basado en Linn, Graue & Sanders, 1991.

Como consecuencia de esto, los puntajes caen dramáticamente en el año 9, seguido de un nuevo incremento constante en los años 10, 11 y 12. Este incremento probablemente no es un resultado de mayores logros de rendimiento. Este es un típico patrón de puntajes de pruebas.

6. Diferentes métodos pueden mostrar diferentes resultados en el logro estudiantil

Las respuestas a preguntas importantes sobre el logro estudiantil pueden variar dependiendo de los datos analizados o cómo son analizados o reportados. Por ejemplo, programas anuales de evaluación (i.e., otoño-a-otoño o primavera-a-primavera) tienden a mostrar un incremento mucho menor en el logro estudiantil que programas de evaluación que usan el ciclo de evaluación otoño-a-primavera (Linn, Dunbar, Harnish & Hastings, 1982). Las diferencias pueden estar causadas por selección de los estudiantes, errores en las escalas de conversión, condiciones de administración, fechas de administración de las pruebas comparadas con las fechas en que se sentaron las normas de referencia de los puntajes de las pruebas, efectos de la práctica y enseñanza de los docentes para la prueba.

Diez sugerencias para los formuladores de políticas

A pesar de estos problemas que afectan a los sistemas de evaluación basados en estándares así como a la mayoría de las mediciones en general, hay ciertas maneras de mejorar la validez, credibilidad e impacto positivo de los sistemas de evaluación a la vez que se minimiza su impacto negativo. Se recomienda que los formuladores de políticas:

1. **Establezcan estándares altos pero asequibles.** Estándares inasequibles llevan al público a creer equivocadamente que no se puede ya mejorar las escuelas. Por otro lado, estándares que no establecen expectativas altas harán que el público pierda la confianza en las escuelas públicas.

2. **Desarrollen estándares y *después* las evaluaciones.** Estudios sobre los niveles de logro del NAEP han mostrado claramente lo errado que resulta intentar imponer niveles de logro o estándares de desempeño a evaluaciones ya existentes. Las revisiones de las pruebas existentes, o las nuevas que se creen, deben medir los estándares cercanamente y reportar con precisión el logro estudiantil.

3. **Incluyan a todos los estudiantes en los programas de evaluación, excepto a aquéllos con las más severas incapacidades. Usar evaluaciones "adaptadas" para los estudiantes que aún no han hecho la transición a programas regulares en inglés o cuyas incapacidades lo requieran**. Esto podría ayudar a asegurar la responsabilización por todos los estudiantes y aumentar la comparabilidad de los resultados de diferentes escuelas y distritos. Reportar puntajes agregados y puntajes separados por subgrupos para

brindar información más exacta y útil sobre el progreso de los estudiantes y las escuelas.

4. **La responsabilización con altas implicancias requiere nuevas evaluaciones de alta calidad cada año que sean comparables a las de años anteriores.** El ahorrar en este sentido probablemente llevará tanto a resultados distorsionados tales como puntajes inflados, como a distorciones en la educación como por ejemplo la enseñanza "estrechada" para las pruebas.

5. **No pongan todo el peso en una sola prueba cuando tomen decisiones importantes sobre los estudiantes y las escuelas (i.e., retenciones, promociones, estados de observación o premios).** Busquen, en cambio, múltiples indicadores del desempeño. Incluyan evaluaciones de desempeño y otros indicadores de logro tales como asistencia, alumnos tomando cursos avanzados, etc.

6. **Pongan más énfasis en las comparaciones de desempeño año a año que a las comparaciones entre escuelas.** Esto toma en cuenta las diferencias en los puntos de partida a la vez que mantiene una expectativa de mejora para todos.

7. **Establezcan metas tanto de corto como de largo plazo para todas las escuelas.** Las metas de corto plazo toman en consideración las diferencias en posiciones iniciales de diferentes escuelas. Las metas de largo plazo permiten expectativas de los mismos estándares altos para todos, incluyendo la expectativa de que las escuelas con menores rendimientos tendrán tasas mayores de crecimiento anual o bienal que las actuales escuelas de mayor rendimiento. Esta combinación dará a las escuelas una oportunidad razonable de mostrar mejoras, al mismo tiempo que impide que se bajen las expectaivas de logro de las escuelas y sus estudiantes.

8. Como en los sondeos de opinión, hay incertidumbre en cualquier sistema de evaluación educacional. Esa **incertidumbre debe reportarse en todos los resultados de la pruebas.**

9. Evaluar no sólo los efectos positivos esperados de las evaluaciones basadas en estándares, **sino también los efectos negativos no esperados de esos sistemas de evaluación.**

10. **Disminuir la brecha de logros significa que debemos brindar a todos los estudiantes los profesores y recursos que necesitan para alcanzar nuestras altas expectaivas.** Esto significa mejorar el sistema educativo como un todo, no sólo realizar más evaluaciones o nuevos sistemas de evaluación.

Referencias

Educación Week. (1997). La calidad cuenta: Una "libreta de calificaciones" sobre las condiciones de la educación pública en los 50 estados. *Un suplemento de Educación Week*. Vol. 16, Enero, 22.

Lerner, L.S. (1998). *Estándares estatales en Ciencias: una evaluación de los estándares de ciencias en 36 estados*. Washington, DC: Fundación Thomas B. Fordham.

Linn, R. L., Dunbar, S. B., Harnisch, D. L., & Hastings, C. N. (1982). La validez de la evaluación y el sistema de reporte del Programa Title I. En E. R. House, S. Mathison, J. Pearsol, & H. Preskill (Eds.), *Revisión Anual de Estudios sobre Evaluación* (Vol. 7, pp. 427-442). Beverly Hills, CA: Publicaciones SAGE.

Linn R. L., Graue, M. E., & Sanders, N. M. (1990). Comparando los resultados estatales y distritales con las normas nacionales: La validez de las declaraciones de que “todo el mundo está sobre el promedio”. *Educational Measurement: Issues and Practice, 9(3), 5-14*.

Olson, L. (1998, abril 15). Una “A” o una “D”: Los *rankings* estatales difieren ampliamente. *Educación Week*, 17, 1, 18.

Raimi, R. A., & Braden, L. S. (1998). *Los estándares estatales en matemáticas: una evaluación de los estándares de ciencias en 46 estados, el Distrito de Columbia y Japón*. Washington, DC: Fundación Thomas B. Fordham.

Reese, C. M., Miller, K. E., Mazzeo, J., & Dossey, J. A. (1997). *La "libreta de calificaciones" del NAEP en matemáticas para la nación y los estados de 1996.* Washington, DC: National Center for Educational Statistics.