



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

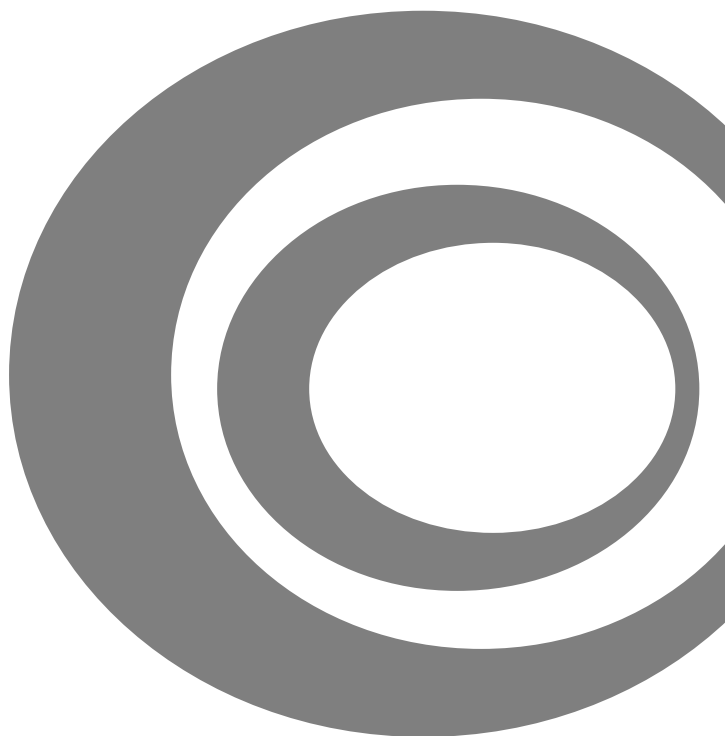
Grupo de Trabajo sobre Estándares y Evaluación

**La evaluación
impulsada por
estándares: Problemas
técnicos y políticos en
la medición del
progreso de la escuela
y los estudiantes**

**Robert L. Linn
Joan L. Herman**



Grupo de Análisis para el Desarrollo



LA EVALUACIÓN IMPULSADA POR ESTÁNDARES: PROBLEMAS TÉCNICOS Y POLÍTICOS EN LA MEDICIÓN DEL PROGRESO DE LA ESCUELA Y LOS ESTUDIANTES

Robert L. Linn y Joan L. Herman
CRESST/ Universidad de California, Los Ángeles

Tomado del *Reporte Técnico del Centro de Estudios sobre Evaluación (CSE) No. 426* CRESST (*Centro Nacional para la Investigación sobre Evaluación, Estándares y Medición del Logro Estudiantil*) / Universidad de California, Los Angeles, con autorización de sus editores.

Resumen Ejecutivo

Los estados de todo el país están estableciendo estándares nuevos y exigentes que definen qué es lo que los estudiantes deben saber y qué es lo que deben ser capaces de hacer. Muchos estados están también diseñando e implementando nuevos sistemas de evaluación, con el fin de ayudar a los estudiantes a cumplir estos estándares y medir su progreso hacia alcanzarlos.

Las evaluaciones juegan un rol fundamental en la reforma liderada por estándares:

- Comunicando las metas que se espera logren los sistemas escolares, escuelas, maestros y estudiantes.
- Proporcionando objetivos para la enseñanza y el aprendizaje, y
- Dando forma al desempeño de educadores y estudiantes.

Unidas a incentivos y/o sanciones apropiadas -- sean externos o auto-dirigidos - las evaluaciones pueden motivar a los estudiantes a aprender mejor, a los maestros a enseñar mejor y a las escuelas a ser más efectivas educacionalmente.

1. ¿Qué hace diferentes a las evaluaciones impulsadas por estándares?

A diferencia de las evaluaciones más tradicionales, las evaluaciones impulsadas por estándares están estrechamente vinculadas al currículo, produciendo una conexión estrecha entre lo que se enseña y lo que se evalúa. A diferencia de las pruebas que referidas a normas, que comparan el desempeño de cada estudiante con el de otros, las evaluaciones impulsadas por estándares incorporan metas de desempeño previamente establecidas. Y, a diferencia de los exámenes de opción múltiple, muchas evaluaciones impulsadas por estándares requieren que los estudiantes demuestren una amplia gama de habilidades para la solución de problemas -- las mismas habilidades que los estudiantes necesitarán para su éxito futuro.

Estas evaluaciones "auténticas" o evaluaciones "de desempeño" suelen hacer que los estudiantes se involucren en problemas de la vida real más que en ejercicios artificiales. No sólo miden la habilidad que tienen los estudiantes para manejar tareas complejas, sino que también sirven de modelo para los maestros, proporcionándoles ejemplos a ser usados en el salón de clases.

Las evaluaciones de desempeño que requieren respuestas extensas deben ser calificadas por jueces expertos, usando guías de calificación claramente especificadas. El desarrollo y la aplicación de dichas guías de calificación brindan a los maestros una oportunidad inusitada de discutir estándares nuevos y expectativas de desempeño.

Examinar las respuestas que dan sus alumnos permite a los maestros comprender los puntos fuertes y las debilidades del aprendizaje de sus pupilos y planear actividades instruccionales apropiadas.

¿Qué es lo que hace sólida a una evaluación? Típicamente, se citan dos criterios principales: validez, el grado hasta el cual se justifican usos e interpretaciones particulares de los resultados de la evaluación; y confiabilidad, el grado hasta el cual las calificaciones están libres de error de medición. Para la evaluación impulsada por estándares, otra clave es el alineamiento -- el grado hasta el cual la evaluación refleja adecuadamente los estándares en los cuales se supone está basada. Una evaluación mal emparejada con un conjunto dado de estándares puede minar el aprendizaje al enfocar la atención en habilidades o conocimientos menos importantes a costa de otros más importantes.

2. Retos para los sistemas de evaluación impulsada por estándares

2.1 Construir un consenso estatal y local

Si nos guiamos por las encuestas de opinión pública, el concepto de que los estudiantes deben alcanzar estándares académicos altos goza de amplio respaldo. Sin embargo, la experiencia muestra que dicho respaldo puede ser frágil. La diversidad de opiniones sobre qué es lo que los estudiantes deben aprender y qué es lo que las escuelas deben enseñar hace imperativo el involucrar al público en el desarrollo de estándares y evaluaciones. Construir un consenso amplio requiere no sólo una serie de audiencias públicas y oportunidades para hacer sugerencias y revisiones, sino un proceso integral que involucre intensamente al público, asegurando que sus preocupaciones sean comprendidas y atendidas.

2.2 Proporcionar estándares fuertes

Lograr un consenso sobre estándares que son vagos y amplios no es ningún reto -- ¿quién estaría en desacuerdo con que todos los estudiantes deben ser capaces de “comunicarse efectivamente?” Pero cuando los estándares están formulados en términos tan generales, son de poca ayuda para los estudiantes que los deben lograr o para los maestros y escuelas que intentan evaluar el progreso estudiantil. La evidencia disponible sugiere que los estándares actuales de muchos estados no son lo suficientemente fuertes como para sustentar una evaluación rigurosa basada en contenidos.

2.3 Alinear los estándares con la evaluación y la enseñanza

Muchos estados y localidades desarrollan estándares y evaluaciones al mismo tiempo, en lugar de seguir la secuencia más lógica: primero los estándares, luego las evaluaciones. De hecho, algunos estados construyen sistemas de evaluación en base a “retazos” de cuantas evaluaciones y pruebas puedan encontrar, sacrificando el “entalle perfecto” que obtendrían si desarrollaran evaluaciones empezando por el principio. Los sistemas que se basan exclusivamente en exámenes de opción múltiple no pueden mostrar qué tan bien se están desempeñando los estudiantes en la gama completa de habilidades y comprensiones que abarcan los estándares.

Por último, el currículo y la instrucción en el aula también deberían estar alineados con los estándares y las evaluaciones. Sin embargo, este alineamiento depende a su vez de la capacidad de los maestros para comprender las expectativas incorporadas en las nuevas evaluaciones para obtener recursos y experiencia para ayudar a sus alumnos a alcanzarlas. Es cuestión de justicia que los estudiantes no sean responsabilizados por metas que no han tenido oportunidades adecuadas de alcanzar.

2.4 Asegurar mediciones precisas

Las evaluaciones de desempeño, que piden a los estudiantes generar una respuesta en lugar de elegir una de una lista, generalmente proporcionan un mejor indicador de habilidades para el pensamiento complejo. Pero calificar evaluaciones como esas requiere más tiempo, generalmente más dinero y un consenso entre jueces sobre la calidad de la respuesta. Para proporcionar un estimado estable sobre la capacidad estudiantil, la mayoría de las evaluaciones que ahora se desarrollan incorpora una variedad amplia de tareas, reflejando así el alcance completo de los estándares. Cuando se mide el desempeño o progreso de una escuela o distrito, en lugar del de un estudiante individual, las evaluaciones pueden también asignar tareas diferentes a muestras diferentes de estudiantes (una práctica conocida como muestra matricial).

2.5 Definir el progreso

El progreso de escuelas, distritos y estados se define típicamente por el desempeño de cohortes sucesivas de estudiantes: ¿Están, por ejemplo, más chicos de 4º grado demostrando su proficiencia en estándares matemáticos este año que el año pasado? La ley federal requiere que los estados definan un “progreso anual adecuado” en términos del desempeño estudiantil en las evaluaciones estatales impulsadas por estándares, que determinen si sus escuelas están logrando dicho progreso, y que señalen como meta una fecha “apropiada” para la cual todos sus estudiantes del Título I¹ se desempeñarán ya sea en el nivel proficiente o en el avanzado. Los estados deben luego establecer una tasa anual de mejoría que sea tanto “sustantiva” como “suficiente” para lograr esa meta.

2.6 Establecer las implicancias (*stakes*)

Lo que las escuelas hacen con los resultados de las evaluaciones -- ya sea simplemente reportándolas, en un extremo del espectro, o condicionando la graduación de los estudiantes, al otro extremo -- puede tener efectos profundos sobre esos estudiantes. Las evaluaciones también pueden utilizarse para responsabilizar a los educadores y a las escuelas por el desempeño de los estudiantes. Los distritos pueden utilizar los resultados como base para premios explícitos (otorgamiento de donaciones en efectivo, por ejemplo) o sanciones (reasignación o destitución de personal, reorganización administrativa).

2.7 Incluir a todos los estudiantes

Los estándares están diseñados para elevar las expectativas que se tienen para todos los estudiantes. El incluir en una evaluación a estudiantes con limitaciones en la lengua inglesa y a estudiantes discapacitados puede requerir una variedad de distintas estrategias para brindarles facilidades, desde la asignación de tiempo extra hasta la provisión de evaluaciones orales o la traducción a otras lenguas. Los estudiantes con problemas de aprendizaje, que históricamente ha constituido el grupo más grande excluido de las evaluaciones, pueden ser capaces de rendirlas, ya sea en parte o por completo, sin facilidades especiales. (Aquéllos con problemas cognitivos severos pueden requerir de un sistema de evaluación distinto).

2.8 Estimar costos

Aunque los costos de las evaluaciones varían considerablemente, aquéllas que requieren respuestas extensas de los estudiantes -- a ser evaluadas por maestros u otros expertos en las disciplinas -- cuestan sustancialmente más que las pruebas de

¹ Título I (*Title I*) es un amplio programa del gobierno federal de los E.E.U.U. 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.

opción múltiple. Administrar una prueba que pueda ser calificada automáticamente puede costar entre \$5 y \$8 por estudiante; las evaluaciones que requieren de una combinación de respuestas cortas y respuestas escritas extensas pueden costar fácilmente dos o tres veces eso. Los estimados para las evaluaciones de desempeño más complejas varían entre \$30 y \$70 por alumno.

2.9 Encarar retos legales

Las evaluaciones tienen que enfrentar retos legales, especialmente cuando sus resultados tienen altas implicancias o costos (sea para que un estudiante se gradúe o para endosar un diploma). También se pueden esperar cuestionamientos cuando las evaluaciones producen consecuencias adversas para grupos históricamente en desventaja: por ejemplo, tasas de fracaso sustancialmente más altas entre estudiantes afro-americanos o hispanos. Esto en si mismo no es evidencia suficiente de la injusticia de una evaluación o de un intento de discriminar. Pero la identificación de las consecuencias adversas puede -- y muchas veces lo hace -- detonar cuestionamientos legales.

Entre los más probables desencadenantes, están:

- El “uso de procesos percibidos como injustos, arbitrarios o caprichosos”.
- La “sugerencia de que se están evaluando actitudes o valores específicos”.
- El “no proporcionar todas las facilidades requeridas por los discapacitados”.
- La evaluación de “conocimientos o habilidades que los examinados no han tenido la oportunidad de aprender”.

La probabilidad de cuestionamientos legales es un argumento en contra de atar tempranamente altas implicancias o costos a los resultados de evaluación. Diseñar un sistema confiable de evaluaciones ligadas a estándares es un proceso complejo que toma tiempo. Asimismo tomará tiempo que los maestros, escuelas y alumnos comprendan las expectativas que levanta un sistema así y que puedan cumplirlas.

2.10 Construir capacidad local

La investigación muestra que la mayoría de los maestros toma con seriedad las evaluaciones de desempeño e incorpora las metas que las subyacen en su práctica de enseñanza. Al mismo tiempo, sin embargo, muchos directores y maestros reportan serias preocupaciones por las demandas que las nuevas evaluaciones exigen de ellos mismos y de sus escuelas. Declaran, en particular, que los maestros necesitan tiempo para familiarizarse con los nuevos estándares, evaluaciones y requerimientos administrativos; para comprender cómo se desarrollan y se califican las nuevas formas de evaluación; para aplicar criterios al evaluar el trabajo de los estudiantes; y para adquirir información y conocimientos pedagógicos suficientes como para permitirles cambiar sus prácticas. Es igualmente importante proporcionarles recursos apropiados y oportunidades suficientes para su desarrollo profesional.

2.11 Diferenciar las evaluaciones

Una evaluación que intenta desempeñar demasiadas funciones -- diagnóstico del estudiante, planeamiento curricular, evaluación de programas, mejoría pedagógica, responsabilización, certificación y comunicación social -- inevitablemente no hará ninguna bien. Por ello, es importante distinguir roles apropiados para diferentes evaluaciones a nivel distrital, escolar y del aula. Un sistema cohesionado asegura que tanto maestros como alumnos comprendan qué es importante aprender y qué tan bien lo están haciendo.

Los maestros usan rutinariamente una amplia variedad de evaluaciones formales e informales para medir el progreso de los estudiantes, dar calificaciones, motivar la atención, proporcionar retroalimentación y adaptar la instrucción a las necesidades del alumno. De manera similar, los alumnos se involucran regularmente en la auto-evaluación, mientras estudian y monitorean su propio progreso. Juntas, todas estas evaluaciones proporcionan a maestros y alumnos la comprensión detallada y la continua retroalimentación necesaria para guiar un aprendizaje continuo y efectivo. Es esencial que estas evaluaciones reflejen los estándares estatales.

Introducción y Antecedentes

Mejorar el desempeño del estudiante requiere una visión clara de aquello que usted desea lograr, un sistema integral de medición para captar el progreso, y un compromiso de actuar según los resultados para realizar cambios apropiado.

Roy Romer, Gobernador de Colorado

Los estados en todo el país están estableciendo estándares nuevos y exigentes que definen qué es lo que los estudiantes deben saber y deben ser capaces de hacer. Muchos estados están también diseñando e implementado nuevos sistemas de evaluación para ayudar a los estudiantes a lograr estos estándares.

Estos sistemas son muy prometedores como apoyo a un mejor desempeño de los estudiantes, pero su efectividad depende de un número de factores. Este documento señala los más importantes, así como algunas de las lecciones aprendidas a lo largo de la última década por los estados y las localidades involucradas en el debate sobre la evaluación.

Los estándares deben ser lo suficientemente específicos como para permitir a todos (estudiantes, padres, educadores, diseñadores de políticas y la opinión pública en general) entender aquello que los alumnos necesitan aprender. También deben ser lo suficientemente precisos como para permitir una apreciación justa y precisa sobre si se han cumplido o no. Aunque los estándares no prescriben un currículo, un libro de texto ni un enfoque pedagógico específico, y se puedan lograr de diversas maneras, los estándares sí deben dejar muy claro qué es lo que se espera de los alumnos.

1. Estándares de Contenido y de Desempeño

Los estados y localidades normalmente distinguen dos tipos de estándares inter relacionados: aquéllos que especifican el contenido (lo que los alumnos deben saber o deben ser capaces de hacer en diferentes momentos de su educación); y aquéllos que especifican el desempeño (qué tan bien deben ser capaces de hacerlo). Lo ideal es que los estándares de desempeño señalen qué evidencia se requiere para demostrar el logro de los estándares de contenido (por ejemplo, un ensayo, una demostración matemática, un experimento científico, un proyecto, un examen) así como la calidad de desempeño que se considerará aceptable (es decir, que amerita una nota aprobatoria o una "A") (Panel Nacional de Metas Educativas, 1993).

Al elevar las expectativas para todos los alumnos, los estándares marcan un primer paso importante para el mejoramiento de la educación. Pero los estándares por sí solos no pueden producir el mejoramiento deseado. Son igualmente instrumentales las especificaciones y materiales curriculares, las guías de recursos, el desarrollo profesional y las evaluaciones. Aunque nuestro enfoque se limita a las evaluaciones, el éxito de una reforma impulsada por estándares requiere un conjunto de cambios sistemáticos a todo lo largo y ancho del sistema educativo (ver las figuras 1 y 2).

Cuadro 1. Ejemplos de Estándares de Contenido**Estándar 4, Geografía, Colorado.**

“Los estudiantes comprenden cómo los procesos económicos, políticos, culturales y sociales interactúan y producen distintos modelos de poblaciones humanas, dependencia recíproca, cooperación y conflicto... En el nivel K-4, lo que los alumnos deben saber y deben ser capaces de hacer incluye... identificar las causas de la migración de los seres humanos (Modelos de Estándares de Contenidos para Geografía, Departamento de Educación de Colorado, adoptado en junio de 1996, rectificado el 1º de Noviembre de 1995).

Estándar de Ciencias, Missouri.

En ciencias, los alumnos de las escuelas públicas de Missouri adquirirán una base sólida que incluye el conocimiento de...propiedades y principios de la fuerza y el movimiento” (Los Estándares “Show–Me”, Departamento de Educación Básica y Secundaria de Missouri, Octubre, 1995

Estándar e Indicador de Logro en Lectura, Grados 6 – 8, Oregón

“Demostrar una comprensión inferencial sobre una variedad de materiales impresos.” Indicador de Logro Asociado, Nivel 8: “ Identificar relaciones, patrones de imágenes o símbolos y sacar conclusiones sobre sus significados” (Objetivos curriculares comunes según nivel de grado, Estándares de Contenido y Desempeño para los Grados 6-8, Departamento de Educación de Oregon, Agosto, 1996)

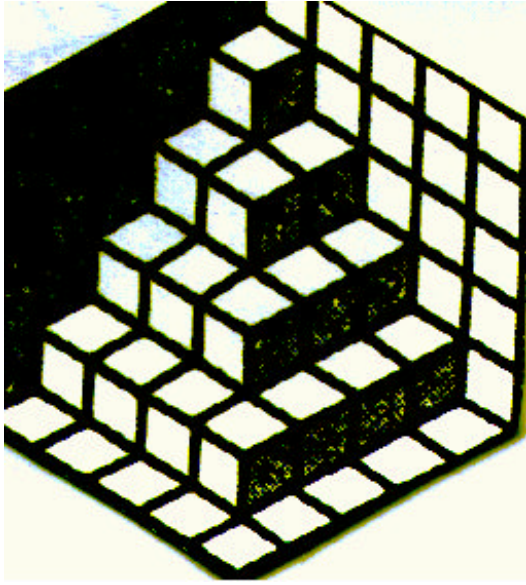
Estándares de Aprendizaje en Historia Norteamericana y Ciencias Sociales, Grado 5, Virginia. Estándar 5.3

“El estudiante describirá la Norte América colonial, poniendo énfasis en ... las principales conexiones económicas y políticas entre las colonias e Inglaterra”. (Estándares de Aprendizaje para las Escuelas Públicas de Virginia, Comisión de Educación, Estado de Virginia; 1995).

Cuadro 2. Ejemplo de una pregunta de opción múltiple y una de desempeño.
(Adaptadas de *“Better Tests Give a Clearer Picture”*, Edmonds School District, Lynnwood, Washington).

ITEM DE OPCIÓN MÚLTIPLE	
1. ¿Cuál de los siguientes números es menor a 50 y mayor a 30?	
1	10
2	60
3	20
4	40

HOJA DE RESPUESTA PARA USO GENERAL			
1	A B C D	10 A B C D	20 A B C D
	ⓧ Ⓝ Ⓜ Ⓟ	ⓧ Ⓝ Ⓜ Ⓟ	ⓧ Ⓝ Ⓜ Ⓟ
2	A B C D	11 A B C D	21 A B C D
	ⓧ Ⓝ Ⓜ Ⓟ	ⓧ Ⓝ Ⓜ Ⓟ	ⓧ Ⓝ Ⓜ Ⓟ
3	A B C D	12 A B C D	22 A B C D
	ⓧ Ⓝ Ⓜ Ⓟ	ⓧ Ⓝ Ⓜ Ⓟ	ⓧ Ⓝ Ⓜ Ⓟ

ITEM DE DESEMPEÑO
<p>¿Cuántos cubos fueron necesarios para construir esta figura? Explique varias maneras diferentes de resolver este problema. ¿Qué patrones ve? Analice los patrones. Use la terminología apropiada.</p>

<hr/> <hr/> <hr/>

2. El rol de la evaluación en las reformas impulsadas por estándares

Las evaluaciones juegan un rol fundamental en las reformas impulsadas por estándares:

- Comunicando los objetivos que se espera logren los sistemas escolares, escuelas, maestros y alumnos.
- Proporcionando metas para la enseñanza y el aprendizaje; y
- Dando forma al desempeño de educadores y alumnos.

Unidas a incentivos y/o sanciones apropiadas -- sean externas o auto-dirigidas - las evaluaciones pueden motivar a los alumnos a aprender mejor, a los maestros a enseñar mejor y a las escuelas a ser más efectivas educacionalmente.

2.1 Las evaluaciones comunican objetivos

Toda evaluación, sea o no impulsada por estándares, revela las expectativas de sus creadores. Los alumnos que buscan adivinar los deseos de sus maestros a menudo hallan más claves en exámenes pasados que en los syllabus de los cursos, las clases mismas o las lecturas asignadas (Madaus, 1988). Con el tiempo, tal como lo describe George Madaus, la “tradicción de exámenes pasados” puede efectivamente definir la currícula, especialmente cuando el desempeño de los alumnos en los exámenes acarrea consecuencias importantes.

2.2 Las evaluaciones proporcionan metas

Las evaluaciones no sólo elucidan estándares: también proporcionan metas de desempeño para la enseñanza. Las evaluaciones enfocan la atención en un conjunto particular de habilidades o conocimientos: aquéllos que deben ser dominados “para lograr el estándar”. Las evaluaciones ofrecen ejemplos operacionales de qué es lo que deben saber o deben ser capaces de hacer los alumnos. También les dicen a los estudiantes qué tan bueno es “suficientemente bueno”, cuando definen diferentes niveles de proficiencia.

2.3 Las evaluaciones dan forma al desempeño

Las reformas impulsadas por estándares se fundamentan en la premisa de que el hacer explícitas las expectativas incitará a un mayor esfuerzo de parte de los maestros y de los alumnos, esfuerzo que las evaluaciones harán se focalice en metas específicas de desempeño. La capacidad de motivar y focalizar el esfuerzo hace de la evaluación un instrumento poderoso en el arsenal didáctico del maestro. Las evaluaciones bien concebidas -- que cubren campos correspondientes a los objetivos y prioridades del curso especificados en el syllabus -- pueden focalizar la atención del alumno en los conocimientos y las habilidades que son considerados como las más importantes de aprender. Evaluaciones mal concebidas pueden incitar a los alumnos a aprender apresuradamente hechos y cifras que serán prontamente olvidados, por el hecho de saber que la simple repetición mecánica asegurará una nota aprobatoria.

La influencia de las pruebas en el comportamiento del alumno no se limita a los exámenes basados en los cursos. La gran cantidad de matrículas en cursos de preparación para exámenes para recibirse de abogados, para certificaciones de las comisiones médicas y para pruebas de admisión para universidades y escuelas para graduados y profesionales, también dan fe de la influencia que tienen los exámenes sobre los estudiantes.

Las evaluaciones pueden dar forma al comportamiento no sólo de los estudiantes sino también de los profesores, quienes a menudo utilizan las pruebas

como modelos para el diseño curricular e instruccional. Los maestros ayudan a los alumnos a prepararse para las pruebas dedicando largos períodos del horario de clases a actividades tipo pruebas, especialmente cuando aumenta la presión por calificaciones altas (ver, por ejemplo, Corbett y Wilson, 1991; Dorr-Bremme & Herman, 1986; Herman & Golan, 1991; Kellaghan & Madaus, 1991; Shepard, 1991; Smith, Edelsky, Draper, Rottenberg & Cherland, 1991). El énfasis exagerado en exámenes de opción múltiple, en particular, ha alentado a los profesores a “adiestrar y matar” (“*drill and kill*”) a sus alumnos con hojas de trabajo sobre habilidades básicas. El resultado es LQEELQO (“Lo que evalúas es lo que obtienes”), otra razón por la cual las evaluaciones correctamente concebidas son tan importantes.

Por último, las evaluaciones locales y estatales también pueden dar forma al comportamiento de sistemas escolares enteros. Elevar las expectativas, particularmente en la forma de pruebas de altas implicancias o costos, puede estimular a las autoridades a buscar los recursos que sus escuelas o distritos necesitan para reforzar los logros de los estudiantes.

3. ¿Qué es lo que hace diferente a una reforma dirigida por estándares?

El rol de las evaluaciones en incrementar la responsabilización y estimular el mejoramiento no es, por supuesto, exclusivo de las reformas impulsadas por estándares. ¿Qué es lo que hace a éstas diferentes de sus contrapartes más tradicionales?

3.1 Conexión estrecha con la currícula

La mayoría de evaluaciones externas en los E.E.U.U., salvo raras excepciones (como los exámenes de Colocación Avanzada y el New York Regents), miden habilidades y logros genéricos intencionalmente desligadas de cualquier currícula, curso o estándares de contenido específicos. La reforma impulsada por estándares, por el contrario, promueve una estrecha conexión entre lo que se enseña y lo que se evalúa. El poder de las evaluaciones de modelar las prácticas docentes, que antes se consideraba como un efecto secundario desafortunado y no intencional, se convierte en algo deseable -- más aun, se ve fortalecido -- al elevarse las implicancias o costos que van ligadas a los estándares.

3.2 Los estudiantes se comparan con estándares de desempeño, no con otros estudiantes

Las evaluaciones impulsadas por estándares comparan los logros de los alumnos con objetivos de desempeño preestablecidos, en lugar de hacerlo con el desempeño de otros alumnos. Se supone que el estándar es absoluto, independiente de la proporción de estudiantes que lo logre. Las pruebas referidas a normas, por el contrario, describen lo que los alumnos pueden hacer en relación con otros alumnos. El hecho de que un alumno alcance el 60avo. percentil en matemáticas, por ejemplo, sólo nos indica que le va tan bien como, o mejor que, al 60% de sus pares, no cuánta habilidad matemática ha llegado a dominar.

El actual enfoque en estándares absolutos refleja un énfasis anterior en “competencias mínimas”, un movimiento de reforma que fue popular en los setentas y ochentas. Entonces como ahora, los reformadores buscaban mejorar la educación haciendo responsables a los profesores y alumnos responsables por el logro de estándares de desempeño, usando para ello pruebas para graduarse de secundaria o para pasar al siguiente grado. Pero, en contraste con los reformadores actuales que dan énfasis a estándares altos y rigurosos, el grupo antes mencionado se enfocaba únicamente en habilidades básicas. Y, a diferencia de las evaluaciones en múltiples niveles impulsadas por estándares, las pruebas de competencia mínima típicamente

empleaban ítems de opción múltiple y sólo daban resultados de aprobado o desaprobado.

3.3 Incorporan nuevas formas de evaluación

Las evaluaciones impulsadas por estándares toman con frecuencia formas nuevas, pidiendo a los alumnos, por ejemplo, que escriban un ensayo, que resuelvan un problema matemático de la vida real o diseñen y conduzcan un experimento científico. A diferencia de las pruebas de opción múltiple, calificadas mediante lectoras ópticas, estas “evaluaciones de desempeño” (también llamadas “alternativas”, “auténticas” o “directas”) las califican seres humanos, quienes examinan el trabajo del alumno y aplican criterios acordados.

Estas evaluaciones captan una gama más extensa de habilidades de pensamiento complejo y de resolución de problemas, habilidades que los alumnos necesitarán para su éxito futuro. El paso hacia la evaluación de desempeño también refleja un nuevo énfasis en el involucramiento constructivo del estudiante en el proceso de aprendizaje. Aunque las pruebas de opción múltiple pueden, cuando están bien diseñadas, hacer algo más que medir simple memorización de hechos específicos o habilidades aisladas, a veces tienden a sobre enfatizar tales habilidades a expensas de un razonamiento más complejo. Esa tendencia puede hacer inadecuadas esas pruebas, particularmente como modelos de instrucción de más alto orden.

Las pruebas de desempeño surgieron como respuesta al pedido de “evaluaciones para las cuales valga la pena enseñar”. Resnick y Resnick (1992) articularon esa demanda en tres “lineamientos para evaluaciones de responsabilización”: (a) “uno obtiene lo que se evalúa”; (b) “no se obtiene lo que no se evalúa”; y (c) “hay que diseñar evaluaciones para las cuales sería deseable enseñaran los educadores”.

Para poder lograr estos propósitos, las evaluaciones deben tener ciertas características, seis de las cuales están resumidas en la Figura 3. La primera “incluir actividades que son valiosas por si mismas”, es la meta de las “evaluaciones auténticas”: involucrar a los alumnos en problemas de la “vida real”, en lugar de en tareas artificiales. Otras características ponen énfasis en la compatibilidad con la enseñanza, el valor y la adecuación de las pruebas para propósitos particulares (por ejemplo, responsabilización o desarrollo profesional).

Cuadro 3. Características deseables en las evaluaciones (Linn y Baker 1996)

- Las tareas de evaluación deben incluir actividades que son valiosas por sí mismas.
- Las evaluaciones deben modelar o ejemplificar la reforma curricular.
- Las actividades evaluativas deben contribuir a la mejora instruccional, enfocándose en objetivos de instrucción consistentes con las metas de las actividades instruccionales.
- Las evaluaciones deben proporcionar un mecanismo para el desarrollo profesional de los docentes.
- Las evaluaciones deben mejorar el aprendizaje involucrando a los alumnos en actividades significativas que son intrínsecamente motivadoras.
- Las evaluaciones deben conducir a una mayor y más apropiada responsabilización.

3.4 Son valiosas para el desarrollo profesional

Las evaluaciones de desempeño son útiles no sólo para medir la habilidad de los alumnos para dominar tareas complejas sino también para modelar esas tareas para los maestros. Estas evaluaciones incorporan los tipos de problemas complejos y abiertos a los que se enfrentarán los alumnos en el mundo real y simultáneamente sirven como ejemplos que los maestros pueden usar en sus aulas.

La necesidad de modelos de evaluación no debería ser una sorpresa. Muchos maestros, si no es casi todos, fueron entrenados en un momento en que la focalización del logro académico estaba puesta en las habilidades básicas más que en los actuales estándares más altos. Muchos de los actuales principios sobre la enseñanza y el aprendizaje efectivos se desarrollaron recién hace poco. La evaluación misma estuvo mucho tiempo descuidada en la formación docente.

Las evaluaciones de desempeño cumplen también otra función en el desarrollo profesional de los maestros. Aquéllas que requieran respuestas extensas tienen que ser calificadas por jueces expertos usando guías de calificación claramente especificadas. La participación de docentes en el desarrollo y aplicación de tales guías no sólo saca partido de sus conocimientos sobre el tema o área, sino también les proporciona una oportunidad poco común para discutir los nuevos estándares y las expectativas que se tienen sobre el desempeño. El examinar las respuestas reales obtenidas ayuda a los maestros a comprender las fortalezas y debilidades del aprendizaje de sus alumnos y planificar actividades instruccionales apropiadas. De hecho, muchos maestros describen su participación en sesiones de calificación bien concebidas como una de las partes más valiosas de su desarrollo profesional.

3.5 Asegurando la calidad de la evaluación en los sistemas impulsados por estándares

¿Qué es lo que hace sólida a una evaluación? Se suele mencionar dos criterios principales: validez, el grado hasta el cual se justifican usos e interpretaciones particulares de los resultados de una evaluación; y confiabilidad, el grado hasta el cual las calificaciones están libres de errores de medición (Figura 4).

Cuadro 4. Criterios para evaluar una evaluación (CRESST, UCLA)

Consecuencias. ¿Hasta qué punto se obtienen las consecuencias positivas buscadas? ¿Cuáles son las consecuencias negativas no – intencionales?

Imparcialidad. ¿La evaluación permite a los estudiantes mostrar lo que saben y pueden hacer, sin importar raza, género o status social?

Transferencia y generalización. ¿Proporcionarán los resultados una generalización correcta sobre los logros de los alumnos?

Complejidad cognitiva. ¿Requiere que los alumnos resuelvan problemas y razonen de manera compleja?

Calidad del contenido. ¿Es el contenido consistente con los mejores y más actualizados conocimientos del tema que se está tratando?

Propiedad lingüística. ¿Permite a los alumnos demostrar lo que saben y lo que son capaces de hacer sin verse desbordados con demandas del lenguaje no exigidas por el contenido?

Sensibilidad instruccional. ¿Produce una instrucción efectiva mejoras en el desempeño?

Importancia curricular. ¿Cuán importantes son los objetivos medidos por las evaluaciones? ¿Miden estándares de contenido importantes?

Cobertura de contenidos. ¿Qué tan extensamente se cubre el abanico completo de elementos centrales de los estándares de contenido?

Significatividad. ¿Encuentran los alumnos que las tareas de evaluación son realistas y que valen la pena?

Factibilidad y costo. ¿Vale la pena el costo o el tiempo que se toma en obtener la información sobre los alumnos?

La investigación reciente ha expandido estos criterios, según muestran los listados de las siguientes páginas. Un listado desarrollado por el Centro Nacional para la Evaluación, Estándares y Medición del Logro Estudiantil (CRESST), pone énfasis en las consecuencias y “justicia” de las evaluaciones². El otro, adoptado por el Consejo Nacional de Maestros de Matemáticas (Consejo Nacional de Maestros de Matemáticas, 1995), incluye “equidad” y “apertura” (Figura 5).

² Ver el web site de CRESST: <http://ww.cse.ucla.edu>

Cuadro 5. Estándares para la Evaluación de Matemáticas Escolares del Consejo Nacional de Maestros de Matemáticas (NCTM) (1995).

El estándar de matemáticas:

“La evaluación debe reflejar las matemáticas que todos los alumnos necesitan saber y ser capaces de hacer”.

El estándar de aprendizaje:

“La evaluación debe incrementar el aprendizaje de las matemáticas”.

El estándar de equidad:

“La evaluación debe promover la equidad”

El estándar de apertura:

“La evaluación debe ser un proceso abierto”

El estándar de inferencias:

“La evaluación debe promover inferencias válidas acerca del aprendizaje de las matemáticas”

El estándar de coherencia:

“ La evaluación debe ser un proceso coherente”.

3.6 Es imperativo el alineamiento con los estándares

¿Cómo difieren -- si lo hacen -- los criterios para juzgar una evaluación impulsada por estándares de la lista de factores que se usan para evaluar cualquier otra prueba? La respuesta, en una sola palabra, es alineamiento: el grado en el cual la evaluación refleja de manera adecuada los estándares en los cuales se supone está basada.

El punto parece obvio, pero considérese la alternativa: una evaluación que se empareja mal con un conjunto dado de estándares puede socavar el aprendizaje al enfocar la atención en habilidades o conocimientos menos importantes a expensas de otros. Si, por ejemplo, los estándares de historia enfatizan la explicación y el pensamiento crítico, entonces una evaluación alineada con estándares debe requerir que los estudiantes piensen críticamente acerca de la información histórica y expliquen tendencias o eventos históricos. Una evaluación “desalineada” que simplemente pide a los alumnos recordar fechas y nombres, frustraría la intención original.

Ninguna evaluación puede captar la serie completa de estándares de contenido de una disciplina dada, pero algunas pueden abarcar más que otras. Como se dijo anteriormente, las pruebas de opción múltiple por sí solas probablemente esquivarán las habilidades y conocimientos más difíciles -- pero no menos importantes -- de evaluar. La discusión en el capítulo siguiente señala la necesidad de una combinación de formatos de evaluación.

Retos para la evaluación impulsada por estándares

Más allá de las preocupaciones técnicas, diseñar un sistema de evaluación efectivo requiere de un gran liderazgo y de planeamiento cuidadoso. Construir un consenso local y estatal y asegurar un nivel de alineamiento suficiente entre evaluación e instrucción son, entre otros retos, las claves de este proceso (Figura 6).

Cuadro 6. Los estándares y las evaluaciones juegan un rol central en los diez puntos del “Llamado a la Acción por la educación americana en el siglo 21” formulado por el Presidente Clinton.

El Presidente propone pruebas nacionales del desempeño individual de los alumnos en lectura en el 4º grado y en matemáticas en 8º grado, ligadas a la Evaluación Nacional de Progreso Educacional y al Tercer Estudio Internacional de Ciencias y Matemáticas. El Presidente también pide a los estados acabar con la promoción social, exigiendo a los alumnos que demuestren lo que han aprendido para pasar de la escuela elemental a la intermedia y de allí a la secundaria, y asegurando que un diploma de secundaria significa algo en realidad.

Por último, el plan del Presidente recomienda la utilización de estándares y evaluaciones para responsabilizar a las escuelas, los administradores y los educadores. Particularmente, anima a los distritos y estados a usar su autoridad bajo el programa reformado Título I para que las escuelas se responsabilicen por la ayuda que reciben, reconstituyendo escuelas que fracasan crónicamente, entre otras medidas.

1. Construir consensos a nivel estatal y local

Si las encuestas de opinión pública indican algo, el concepto de que a los alumnos debe exigírseles estándares académicos altos goza de amplio respaldo. Sin embargo, la experiencia demuestra que dicho respaldo puede ser frágil. “El consenso se quiebra...al ir más allá de esta creencia en la necesidad de estándares y evaluación, hacia cuestiones sobre cuáles deben ser esos estándares y cómo debe enseñarse y evaluar a los alumnos” (McDonell, 1996). Por ejemplo, la cálida recepción que tuvieron los Estándares Curriculares y de Evaluación del Consejo Nacional de Profesores de Matemáticas (NCTM, 1989) en el ámbito nacional, no se ha extendido a los estándares en otras áreas de contenido como la historia y la lengua (ver, por ejemplo, Diegmüller, 1994). De hecho, las iniciativas para establecer estándares en estas áreas han desencadenado feroces debates.

Esta divergencia no es muy sorprendente. La misma especificidad que hace de las evaluaciones y los estándares objetivos educativos valiosos -- al definir aquello que es importante que los estudiantes sepan -- los hace también blanco de críticas. Más aun, “los estándares implican mucho más que determinaciones sobre qué conocimiento es de mayor valor; implican también diferencias sociales y culturales, y sirven frecuentemente como símbolos y sustitutos de esas diferencias” (Cremin, 1989). La cancelación del Sistema de Evaluación del Aprendizaje de California (CLAS), por ejemplo, fue en gran medida el resultado de una oposición organizada a (1) los énfasis en contenidos que brindaban poca atención a habilidades básicas como fonética, ortografía y aritmética; y (2) los ejercicios de evaluación a los que los opositores acusaban de “promover valores inapropiados como la violencia y el cuestionamiento de la autoridad”(McDonell, 1996).

Los opositores activos del CLAS, tanto como los críticos de los estándares y las evaluaciones de otros estados, representan una minoría relativamente pequeña, aunque muy verbal. Las cuestiones que estos críticos levantan, sin embargo, pueden reflejar preocupaciones más generales: encuestas recientes sobre la enseñanza de las matemáticas y la escritura señalan diferencias fundamentales entre los valores curriculares de los reformadores educativos y grandes segmentos del público (McDonell, 1996).

La diversidad de opiniones acerca de qué es lo que los alumnos deben aprender y las escuelas enseñar, hace imperativo involucrar al público en el desarrollo de estándares y evaluaciones. Mientras que los reformadores educativos pueden creer firmemente en las teorías “constructivistas” del aprendizaje, también tiene que considerarse otras habilidades más tradicionales como ortografía, fonética, tablas de multiplicar, conocimiento de fechas y lugares históricos. El ignorar los deseos de grandes segmentos del público pondrá en peligro toda la empresa.

Construir un consenso amplio requiere no sólo una serie de audiencias públicas y oportunidades para proponer y revisar, sino un proceso integral que involucre a fondo al público, asegurando que sus preocupaciones sean comprendidas y tomadas en cuenta.

Las experiencias de California, Kentucky y Carolina del Norte destacan la necesidad de un liderazgo político fuerte, especialmente, como McDonell (1996) ha demostrado, cuando la introducción de reformas impulsadas por estándares requiere cambios mayores:

Aproximaciones fundamentalmente diferentes a la enseñanza y la evaluación necesitan de voceros claros que crean firmemente en las ideas y que puedan persuadir a los padres y al público en general de que estas estrategias producirán ganancias positivas, tanto para los alumnos individualmente como para el estado como un todo. Ese apoyo tiene que venir de personas destacadas y de personas a quienes el público siente que puede responsabilizar por los resultados. Por esa razón, el apoyo debe venir de personas que ocupen cargos elegidos, no sólo de educadores profesionales o de funcionarios no electos del sector educativo.

2. ¿Proveen los estándares un fundamento sólido para la evaluación y el alineamiento?

Conseguir un consenso sobre estándares que son amplios y vagos no es ningún reto. ¿Quién no estaría de acuerdo en que todos los alumnos deben ser capaces de “comunicarse efectivamente?” Pero cuando los estándares se establecen en términos tan generales, son de poca ayuda para los estudiantes que deben alcanzarlos o para los maestros y escuelas que intentan evaluar el progreso de sus alumnos.

Considérese, por ejemplo, un estándar que exige que los alumnos sean capaces de “comprender ideas y documentos dentro de contextos históricos”. Una variedad de preguntas de opción múltiple, respuestas cortas o ensayos podrían ser herramientas de evaluación apropiadas. Pero, ¿qué ideas o documentos deben comprender los alumnos? ¿En qué contexto? Y, ¿qué constituye una evidencia adecuada de comprensión?

Ya que el estándar mismo proporciona pocas claves, la respuesta debe venir de aquéllos que desarrollan la evaluación, idealmente, vía una discusión con educadores, estudiantes, padres y el público en general. Una vez que se ha desarrollado una evaluación adecuada, se deben generar algunos criterios para la calificación de las

respuestas de los alumnos. Incluso los estándares explícitos dejan bastante espacio para la especificación de tareas de evaluación y para distinguir los niveles de desempeño de los alumnos.

Para que sean efectivos, los estándares deben “estar escritos en un lenguaje claro, explícito... firmemente enraizados en el contenido del área temática, y ... suficientemente detallados como para proporcionar una guía importante para los maestros, diseñadores de currícula y evaluaciones, padres, alumnos, y otros que los usarán” (Federación Norteamericana de Maestros, 1996). Pero la evidencia disponible sugiere que los estándares actuales de muchos estados no son lo suficientemente sólidos como para sostener una evaluación rigurosa basada en contenidos. La falta de alineamiento amenaza el éxito de los esfuerzos estatales en materia de reformas.

3. ¿Están los estándares alineados con la evaluación y la enseñanza?

Es más fácil hablar de alineamiento que hacerlo realidad. La dificultad se deriva en parte de las decisiones de los estados y localidades de desarrollar estándares y evaluaciones al mismo tiempo, en lugar de seguir la secuencia más lógica: primero los estándares y luego, las evaluaciones. Esta decisión, algunas veces resultado de presiones de presupuesto y tiempo, hace más complicado el alineamiento. De hecho, algunos estados arman sistemas echando mano a cuanto evaluación esté disponible, sacrificando “la hechura a la medida” que obtendrían si desarrollaran evaluaciones partiendo desde cero. (Hay, por supuesto, excepciones notables).

El resultado es, a menudo, una correspondencia superficial entre estándares y evaluaciones. Ambos pueden cubrir los mismos temas, pero no llegan a alinearse en otros sentidos (Webb, 1997). ¿Reflejan las evaluaciones y los estándares los conceptos centrales y los temas perdurables de la disciplina? ¿Requieren las tareas de evaluación los tipos de pensamiento complejo y las capacidades para resolver problemas especificados por los estándares? ¿Son los tipos de situaciones problemáticas similares e igualmente auténticas?

Pocos sistemas pueden superar esta prueba. Aunque casi todos los estados y muchos distritos han desarrollado (o están desarrollando) estándares altos para el desempeño de los alumnos (Federación Norteamericana de Maestros, 1996), muchos programas de evaluación aún se basan exclusivamente en pruebas estandarizadas de opción múltiple (Bond, Braskamp, & Roeber 1996). Los datos que producen estos programas pueden decir a las escuelas, a los padres o al público qué tan bien se están desempeñando los alumnos en sólo algunos aspectos de los estándares, no en todo el rango de habilidades y comprensiones cubierto por los estándares.

En última instancia, la currícula y la instrucción en el aula deben estar alineadas con los estándares y las evaluaciones. Si este alineamiento no está presente, los estudiantes no pueden adquirir el conocimiento y las habilidades que necesitan para lograr los estándares. Pero este alineamiento depende a su vez de la habilidad de los profesores para comprender y obtener los recursos y las capacidades que les permitan a sus alumnos cumplir las expectativas incorporadas en las nuevas evaluaciones. La equidad demanda que los alumnos no sean responsabilizados por objetivos que no han tenido oportunidad adecuada de alcanzar.

4. Definir niveles distintos de desempeño

Los estándares de contenido y los de desempeño expresan lo que los alumnos deben saber y deben ser capaces de hacer para demostrar que han logrado el aprendizaje que necesitarán para su éxito futuro. Pero el deseo de mapear el progreso

de los alumnos con más detalle ha llevado a muchos estados a definir no sólo “lo que es suficientemente bueno”, sino también distintos niveles de desempeño.

La ley de Mejoramiento de las Escuelas Norteamericanas de 1994 (Ley Pública 102-382) exige que los estados adopten tres niveles de desempeño: “proficiente”, “avanzado” y “parcialmente proficiente”. El nivel proficiente indica que un alumno ha logrado los estándares de contenido; el nivel avanzado indica que un alumno los ha superado. Los alumnos de menor desempeño son designados como parcialmente proficientes. Los requerimientos de responsabilización del Título I buscan ayudar a que todos los estudiantes logren la proficiencia o demuestren anualmente un progreso adecuado hacia ella.

Los niveles de desempeño de los alumnos se establecen generalmente mediante un proceso de consenso público. Se reúne a paneles de profesores, padres, estudiantes, y miembros de la comunidad empresarial para que revisen trabajos de alumnos que muestran diversidad de habilidades, y para que determinen en qué nivel debe ser clasificado el trabajo. Los miembros de los paneles, por lo general, emiten juicios individualmente, discuten sus razones y, luego -- ayudados por programas estadísticos que convierten sus juicios en calificaciones propuestas -- consideran las implicancias a la luz de los datos reales sobre el desempeño (Wiley, en prensa). Sobre la base de este proceso, se establece una serie de "puntajes de corte", permitiendo que el desempeño del alumno en las evaluaciones se convierta en un nivel de proficiencia.

5. Asegurar mediciones precisas en sistemas impulsados por estándares

Si bien pedirle a los alumnos que generen una respuesta -- en lugar de elegir una de una lista -- puede proporcionar una mejor medida de habilidades de pensamiento complejo, estas evaluaciones de desempeño presentan retos especiales para aquéllos que las califican. El problema no es la introducción del juicio “subjetivo” humano; también son seres humanos quienes están involucrados en la creación de las pruebas de opción múltiple, si bien no en su calificación. Pero calificar evaluaciones de desempeño requiere más tiempo, usualmente más dinero y, a menudo, lo más complicado: lograr un consenso entre jueces sobre la calidad de la respuesta.

Aparte de asegurar un consenso sobre las calificaciones, ¿cómo podemos estar seguros de que las calificaciones son justas y precisas? La respuesta depende, en parte, del diseño de las evaluaciones³. Deben haber ítems suficientes como para obtener un estimado estable de la capacidad de los alumnos. La mayoría de las evaluaciones que ahora se llevan a cabo incorporan por lo tanto una gran variedad de tareas, reflejando el alcance completo de los estándares. (Administrar algunas de estas tareas sólo requiere un corto período de tiempo; otras requieren un período considerablemente mayor.)

Al medir el desempeño o progreso de una escuela o distrito en lugar del de un alumno en particular, las evaluaciones pueden asignar tareas diferentes a muestras diferentes de alumnos. Esta aproximación, conocida como muestreo matricial, asegura una cobertura completa, a la vez que minimiza el tiempo que cada alumno necesita

³ La evidencia disponible sugiere que se requieren entre 5 y 20 tareas para obtener una estimación confiable. Ver, por ejemplo, Baker, 1994; Dunbar, Koretz y Hoover, 1991; Linn, Burton, DeStefano y Hanson 1995; Shavelson, Baxter y Pine, 1991; Shavelson, Baxter y Gao, 1993; Shavelson, Mayberry, Li y Webb, 1990.

para rendir la prueba. (Nótese que, generalmente, el muestreo matricial no es apropiado para evaluar el desempeño de los individuos.)

6. ¿Qué significa progresar?

Las evaluaciones son diseñadas para medir el progreso educacional de los alumnos, ya sea individualmente o como miembros de grupos más grandes. El progreso de las escuelas, los distritos y los estados se suele definir según el desempeño de cohortes sucesivas de alumnos: Por ejemplo, ¿están este año más alumnos de cuarto grado demostrando habilidad en los estándares de matemáticas que el año pasado? Los cambios en la distribución de otras calificaciones son igualmente importantes: ¿Qué proporción de alumnos ha pasado de "proficiente" a "avanzado"? ¿Cuántos permanecen "parcialmente proficientes"?

¿Qué podría considerarse un progreso razonable? Los programas federales del Título I exigen que los estados definan un "progreso anual adecuado" en términos del desempeño de los alumnos en la evaluación impulsada por estándares y que determinen si sus escuelas están logrando dicho progreso. Para poder hacerlo, los estados deben fijar una fecha "apropiada" en la cual todos sus alumnos del Título I se estarán desempeñando ya sea en el nivel proficiente o en el avanzado. Los estados deben luego establecer una tasa anual de mejoría en el desempeño que sea tanto "sustancial" como "suficiente" para la consecución de dicho objetivo (Ley Pública 102-382).

Con el fin de promover la responsabilización por los sub-grupos de alumnos que están en mayor riesgo, la ley exige que los resultados sean desagregados y reportados por separado según estado, distrito o nivel escolar, y según "género, grupo étnico y racial, status de proficiencia en inglés, estado migratorio, alumnos con discapacidades en comparación con aquéllos que no las tienen, y alumnos en desventaja económica en comparación con aquéllos que no están en desventaja económica" (Ley Pública 102-382).

El sistema de responsabilización de Kentucky representa una aproximación a estas exigencias (Departamento de Educación de Kentucky, 1994). Las cuatro categorías de desempeño del estado -- "distinguido", "proficiente", "aprendiz" y "novato" -- corresponden gruesamente a las categorías especificadas por el IASA (avanzado, proficiente, parcialmente proficiente, y por debajo de parcialmente proficiente). Las categorías conllevan puntajes de 140, 100, 40 y 0, respectivamente.

Kentucky busca ayudar a que cada escuela alcance un puntaje promedio de 100, -- calificación posible si todos los alumnos se desempeñaran en el nivel proficiente, o, digamos, si el 50 % fueron proficientes, 30 % fueron distinguidos y el 20 % fueron aprendices ($.5 \times 100 + .3 \times 140 + .2 \times 40 = 100$). Bajo esta fórmula, las escuelas pueden lograr el progreso no sólo incrementando el porcentaje de alumnos "distinguidos" sino también reduciendo la proporción de "novatos". El sistema de Kentucky ilustra la variedad de factores que pueden modelar las definiciones de otros estados sobre el progreso anual adecuado.

Cualquiera sea la definición que elija un estado, los estándares y evaluaciones de desempeño deben ser comparables de un año al otro. En otras palabras, una calificación de proficiente en 1999 debe necesariamente representar el mismo nivel de logro (en una área dada) que una calificación proficiente en 1998 o 1997. Mantener esta consistencia requiere prestar considerable atención al diseño técnico de las evaluaciones: el número de tareas, el muestreo de tareas, la reutilización de tareas y cuestiones similares.

7. Responsabilización e implicancias o costos: ¿cómo se reportan y utilizan las calificaciones?

Lo que las escuelas hagan con los resultados de la evaluación puede tener efectos profundos sobre los alumnos. Como mínimo, al reportar a los padres del alumno sobre su desempeño, les hacen prestar atención al progreso educacional del niño (o a su falta de progreso). Algunas escuelas atan implicancias más altas a los resultados de la evaluación, exigiendo por ejemplo, trabajo remedial a los alumnos que no logran un estándar específico. En última instancia, la graduación o promoción de un alumno de un nivel a otro puede depender de su desempeño. (Algunas escuelas también registran los resultados de la evaluación del alumno en un “diploma endosado”).

Pueden aplicarse implicancias o costos igualmente altos a los educadores. Por ejemplo, Kentucky publica los resultados de las evaluaciones de colegios completos, en lugar de los de alumnos individuales. El simple hecho de publicar esos resultados en el periódico puede incrementar la responsabilización de los educadores. Los cambios en el desempeño de una escuela también pueden ser utilizados como base para recompensas más explícitas (dinero en efectivo, por ejemplo) o sanciones (reasignación o despido del personal, reorganizaciones administrativas).

Como se dijo anteriormente, las escuelas que reciben fondos del Título I deben demostrar un “progreso anual adecuado” en el desempeño de sus alumnos (Ley Pública 102-382). Las escuelas que no lo logran durante dos años consecutivos, recibirán asistencia técnica. Aquéllas que logran más que un progreso anual adecuado durante tres años consecutivos serán designadas “escuelas distinguidas”.

8. ¿Qué significa *todos* los alumnos?

Los estándares son diseñados para elevar las expectativas que se tienen sobre todos los alumnos. El excluir a grandes grupos de estudiantes de las evaluaciones estatales o distritales (debido a discapacidades o a barreras de lenguaje), no se considera ya aceptable.

Incluir a todos los estudiantes en una evaluación puede requerir diferentes estrategias. Para algunos grupos antes excluidos se necesita poca adaptación, más allá del compromiso con su inclusión. Algunos alumnos pueden necesitar un tiempo adicional para completar una evaluación. (Cuando la velocidad de la respuesta no es una consideración relevante, los límites de tiempo podrían relajarse para todos los alumnos.)

Para acomodar a alumnos con inglés limitado se puede ofrecer pruebas en otros idiomas, permitiendo, por ejemplo, que los hispanohablantes demuestren su proficiencia en matemáticas. Sin embargo, esta aproximación presenta varios retos. Primero, si bien hay docenas de idiomas presentes en las aulas norteamericanas, son pocas lo suficientemente comunes como para que el desarrollo de evaluaciones alternativas sea viable. (En muchos estados, el único idioma con un gran número de hablantes nativos, aparte del inglés, es el español.) Segundo, muchos alumnos que son proficientes oralmente en una primera lengua que no es el inglés pueden no haber recibido instrucción formal en ese idioma y podrían no ser capaces, por lo tanto, de rendir una prueba escrita en su lengua materna. Para estos estudiantes puede ser necesario dar pruebas orales.

Los estudiantes discapacitados pueden necesitar facilidades especiales. Aquéllos que tienen impedimentos visuales por ejemplo, pueden necesitar evaluaciones impresas en caracteres más grandes o en Braille. Algunos estudiantes pueden necesitar ayuda para registrar sus respuestas.

Aquéllos con problemas de aprendizaje constituyen el grupo más grande históricamente excluido de las evaluaciones. Muchos de esos estudiantes, que reciben planes de educación individual, pueden ser capaces de completar evaluaciones, en parte o por completo, sin facilidades especiales. Otros pueden necesitar evaluaciones más cortas, más tiempo para cumplir las tareas, instrucciones orales o respuestas orales (ver por ejemplo, NCE, 1996). Una pequeña fracción (quizás el 0.5% de todos los estudiantes) -- aquélla con impedimentos cognitivos severos -- puede requerir un sistema de evaluaciones separado, dictado por sus IEPs.

9. ¿Qué hay de los costos?

¿Cuánto cuestan las evaluaciones impulsadas por estándares? Es difícil obtener estimados confiables, en parte porque muchos costos asociados a la evaluación como el tiempo que invierten los maestros en prepararla, administrarla, y calificarla, están usualmente absorbidos por las operaciones normales de las escuelas y no costeados en un presupuesto aparte. Los costos de las evaluaciones varían considerablemente, dependiendo del número y la extensión de las respuestas a ser calificadas, el número de jueces o calificadores, la cantidad de áreas de contenido evaluadas, la cantidad y naturaleza de los reportes a elaborarse y la inclusión de "evaluaciones de práctica" y otros materiales de preparación (si los hubiera).

Está claro, sin embargo, que las evaluaciones que requieren respuestas extensas de parte de los estudiantes -- a ser calificadas por maestros u otros expertos en las materias -- suelen costar más que las pruebas de opción múltiple, las cuales pueden calificarse mecánicamente. Administrar una prueba calificable mecánicamente puede costar entre \$5 y \$8 por estudiante, variando según el volumen de las pruebas y la diversidad de servicios de calificación solicitados. (Ese precio cubre normalmente reportes con calificaciones individuales de los estudiantes, reportes de salones y reportes de centros escolares en cinco o más áreas de contenido, así como sub-calificaciones en algunas áreas de contenido). A menudo, las escuelas abaratan costos reutilizando cuadernillos y adquiriendo solamente hojas de respuestas y servicios de calificación después del primer año de administración (Linn, 1995).

Las evaluaciones que requieren una combinación de respuestas cortas y respuestas extensas escritas pueden fácilmente costar dos o tres veces más que las pruebas calificables mecánicamente. Las pruebas del Proyecto de Nuevos Estándares que ofrece la Harcourt Brace Educational Measurement, por ejemplo, cuestan aproximadamente \$22 por estudiante (incluyendo cuadernillos de evaluación, servicios básicos de calificación y un paquete estándar de reportes para evaluaciones en matemáticas o inglés / lenguaje)⁴.

Ninguno de los anteriores estimados incluye los costos operacionales incurridos por las escuelas, distritos o estados. Y los costos de evaluaciones de desempeño más elaboradas -- que abarcan, por ejemplo, la ejecución práctica de tareas en ciencias -- son sustancialmente más altos, con estimados que van de \$30 a \$70 por alumno (McDonnell, 1994). En comparación, las pruebas de Colocación Avanzada en un área

⁴ Harcourt Brace Educational Measurement; **Catálogo: Pruebas y Servicios Relacionados**. San Antonio, 1997. Los estados y distritos asociados al Proyecto Nuevos Estándares actualmente reciben un descuento en el costo por estudiante.

específica cuestan \$ 73 por estudiante, de los cuales \$7 normalmente retornan a la escuela. La mayoría de estas pruebas incluyen tanto una sección de opción múltiple como una sección que exige respuestas extensas de los estudiantes.

10. Aspectos legales y certificación estudiantil con implicancias o costos altos

Las evaluaciones pueden enfrentar una variedad de retos legales. Estos retos tienen mayores posibilidades de surgir cuando se atan altas implicancias o costos a los resultados de las evaluaciones, tales como serían el graduar o no a un estudiante o el endosar o no un diploma.

También se pueden esperar cuestionamientos legales cuando las evaluaciones producen un impacto adverso en grupos históricamente en desventaja: por ejemplo, tasas de fracaso sustancialmente más altas para estudiantes afro-americanos o hispanos. Esta evidencia por sí sola no establece la inequidad de una evaluación ni un intento de discriminación. Pero la identificación de un impacto adverso puede -- y a menudo lo hace -- desencadenar un cuestionamiento legal (Phillips, 1995).

Entre los otros desencadenantes más probables, (según Susan Phillips, abogada y experta en medición, 1995), están:

- El “uso de procesos percibidos como injustos, arbitrarios o caprichosos”;
- La “sugerencia de que se están evaluando actitudes y valores específicos”;
- El “no proporcionar todas las facilidades requeridas por los discapacitados”; y
- La evaluación de “conocimientos o habilidades que los examinados no han tenido oportunidad de aprender.” (p.380).

Los dos últimos retos -- dar facilidades a los discapacitados y asegurar una adecuada oportunidad de aprender -- han demostrado ser los más difíciles. El Acta de los Norteamericanos con Discapacidades de 1990 (Ley Pública 101-336), requiere que se proporcione facilidades razonables a los estudiantes discapacitados. “Las cortes han indicado claramente que se debe compensar con facilidades razonables aspectos de la discapacidad que están al margen de la habilidad que está siendo medida, pero que los administradores de pruebas no están obligados a cambiar la habilidad que está siendo medida con el fin de acomodar al examinado discapacitado” (Phillips, 1995). Pero determinar cuáles aspectos de una discapacidad inciden sobre una habilidad que está siendo medida y qué cambios alterarían la naturaleza de esa habilidad, no es una tarea fácil.

Se han suscitado argumentos con respecto a la “oportunidad de aprender” (“Opportunity To Learn” - OTL) en anteriores casos en los tribunales (incluyendo a Debra P.vs. Turlington, un caso de Florida que cuestiona el requerimiento estatal de competencia mínima)⁵. Este tipo de argumentos también tiene probabilidades de aparecer en cualquier cuestionamiento legal provocado por la evidencia de un impacto adverso: diferencias raciales en los resultados de una evaluación pueden reflejar disparidades en las oportunidades de aprender de los estudiantes.

El debate sobre OTL dio lugar eventualmente a la inclusión de estándares voluntarios espontáneos sobre “oportunidades de aprender” en la ley Metas 2000: Eduquemos a Norte América de 1994 (Ley Pública 103-227). Sus proponentes sostenían que era injusto responsabilizar a los alumnos por el logro de objetivos de desempeño sin darles la instrucción para ello. Sus críticos argumentan que estándares

⁵ Debra P. Turlington, 474 F. Supp. 244 (M.D. Fla. 1979), 644 F. 2d 397 (5t Cir. 1981); 564 F. Supp. 177 (M.D. Fla. 1983), 730 F. 2d. 1405 (11h Cir. 1984).

de OTL restringirían [la variedad de] prácticas locales. La imposición de estos estándares sólo parece será posible mediante nuevos procesos judiciales.

La probabilidad de retos legales constituye un argumento en contra de atar implicancias o costos altos a los resultados de las evaluaciones demasiado pronto. Diseñar un sistema confiable de evaluación impulsada por estándares es un proceso complejo, que requiere mucho tiempo. Tomará otro tanto tiempo el que los maestros, las escuelas y los estudiantes entiendan las expectativas que genera este sistema y que los alcancen.

11. Apoyo y retos para la construcción de capacidades locales

La reforma impulsada por estándares requiere mucho más que la adopción de metas o evaluaciones. Un cambio sistémico de este tipo incluye recursos instruccionales, desarrollo profesional y práctica en el aula.

La introducción de evaluaciones impulsadas por estándares puede, sin embargo, servir como catalizador para otras reformas. Investigaciones en Vermont, Maryland, Arizona, Carolina del Norte y Kentucky han mostrado que la mayoría de profesores toman muy en serio las evaluaciones de desempeño e incorporan los objetivos subyacentes en su enseñanza (ver, por ejemplo, Koretz, Mitchell, Barron, & Keith, 1996). (Ver Cuadros 7 y 8.)

Cuadro 7. Los programas de evaluación estudiantil de Maryland

El Departamento de Educación y el Consejo de Educación de Maryland han desarrollado varias evaluaciones para estudiantes, incluyendo el Programa de Evaluación del Desempeño de las Escuelas de Maryland (MSPAP), el cual evalúa a los estudiantes de tercer, quinto y octavo grados en seis áreas académicas principales; y el Programa de Evaluación Funcional de Maryland (MFTP), que certifica el dominio que tiene el estudiante de las habilidades básicas para graduarse de secundaria. Se proporciona a los distritos una prueba normativa, disponible comercialmente, en lectura, lenguaje y matemáticas, con el fin de situar el desempeño distrital dentro de una perspectiva nacional.

También se está desarrollando un nuevo programa de evaluación secundaria destinado a reemplazar al MFTP y un Programa de Evaluación de Dominio de la Independencia para estudiantes severamente discapacitados en educación especial y educación primaria, para perfilar sus fortalezas y debilidades.

El Programa de Evaluación del Desempeño de las Escuelas de Maryland

El MSPAP surgió de un reporte realizado en 1989 por la Comisión Sondheim de Maryland, la cual hizo un llamamiento a favor de incrementar la responsabilización y el desempeño en las escuelas del estado. Administrado por primera vez en 1991, el MSPAP proporciona información sobre el desempeño escolar en lectura, escritura, manejo del lenguaje, matemáticas, ciencias y estudios sociales.

La evaluación se hace a una prueba matricial, a fin de reducir el tiempo que toma. Los estudiantes son evaluados sólo en porciones de cada área de contenido. Se exige que los distritos escolares administren la Prueba de Habilidades Básicas de California (CTBS/5) al menos a una pequeña muestra de sus alumnos.

Participación del público y de los profesores en Maryland

El MSPAP cuenta con el apoyo del público porque éste ha estado involucrado todo el tiempo. Padres, líderes empresariales y legisladores locales aportaron al establecimiento de estándares de contenido y desempeño para el MFTP y el MSPAP y en la selección del CTBS/5 para el programa de evaluaciones normativas, y continuarán proporcionando insumos para los estándares de contenido y desempeño y para el diseño del nuevo programa de evaluaciones de secundaria.

La participación de los maestros es la piedra angular del continuado éxito del MSPAP. Los profesores desarrollan las tareas de evaluación del MSPAP siguiendo las especificaciones de diseño estatales, califican las pruebas del MSPAP en cuatro centros regionales administrados por un contratista externo, y ayudaron a establecer los estándares de contenido y desempeño del MSPAP. Los maestros y otros educadores locales han estado involucrados en desarrollar estándares de contenido para las escuelas secundarias, en la fase del diseño del nuevo programa de evaluación de secundaria, y participarán en el desarrollo y calificación de las pruebas y en el establecimiento de estándares de desempeño para los exámenes de secundaria.

Información y utilización de resultados

Las calificaciones del MFTP, del MSPAP y del CTBS/5 están disponibles en la forma de "libretas de calificación" para la escuela y para el sistema escolar, llamadas "libretas de notas del desempeño escolar de Maryland". Las calificaciones de las

pruebas y otra información (por ejemplo, tasas de abandono y asistencia) identifican a las escuelas en declive, a las de pobre desempeño y a aquellas que están mejorando. Las escuelas en declive pueden ser elegibles para una reconstitución y emprenden un proceso riguroso de mejoramiento escolar. Algunos sistemas escolares locales ponen en alerta a otras escuelas en declive. Las escuelas que muestran las tasas más altas de mejoramiento reciben un Premio de Reconocimiento al Desempeño Escolar, en la forma de fondos para financiar sus esfuerzos continuados por mejorar.

En el año académico de 1995-1996, todos excepto cinco de los veinticuatro sistemas escolares de Maryland lograron calificaciones más altas que el año anterior. Los sistemas escolares anunciaron mejorías en 15 de las 18 áreas de contenido. Los últimos seis años también han dejado ver mejorías continuas en las tasas de abandono y asistencia estudiantil.

Cuadro 8. Las Evaluaciones del Nuevo Siglo de Carolina del Norte

En 1993, la Asamblea General de Carolina del Norte formó una comisión para desarrollar un sistema de evaluación justo y válido que mediría el conocimiento de los estudiantes en términos del mundo real y proporcionaría una mayor retroalimentación a las escuelas y maestros. El resultado: La Evaluación del Próximo Siglo para Carolina del Norte. La propuesta de la comisión se basa en cuatro principios:

1. Un buen sistema de responsabilización hace más que simplemente auditar el desempeño estudiantil: lo mejora.
2. Si se va a dar una reforma genuina, la evaluación debe ser creíble y abierta.
3. El lema de un sistema de evaluaciones efectivo debe ser "Confía pero verifica".
4. Un sistema de evaluaciones efectivo debe construir capacidades locales para realizar una evaluación de calidad, en lugar de evaluar desde fuera una vez al año.

La Evaluación del Próximo Siglo utilizará pruebas estandarizadas, tareas basadas en el desempeño y portafolios o colecciones del trabajo de los estudiantes para promover la responsabilización y proporcionar información diagnóstica sobre el logro de alumnos individuales. La evaluación propuesta para el cuarto, octavo, décimo y doceavo grado será complementada con una evaluación comprensiva que se rendirá entre el décimo grado y el doceavo, así como con un proyecto para graduarse que exige lectura extensiva, escritura, y una presentación oral.

El nuevo sistema de evaluación amplía la actual evaluación estandarizada de Carolina del Norte "Responsabilización en Capacidades Básicas con Control Local" (*Accountability in the Basics with Local Control – ABCs*). Los componentes clave del nuevo sistema incluyen:

- Tareas de desempeño estatales que requieren la aplicación efectiva de los estándares estatales;
- Sistemas de supervisión variados que aseguren que los estándares locales sean consistentes con los estándares estatales y que la calificación local sea confiable; y
- Un portafolio con el trabajo del estudiante (una colección de evidencias sobre sus logros que incluya sus calificaciones en pruebas estatales, en su trabajo local y los resultados de las tareas de desempeño estatales), a ser calificado localmente contra los estándares estatales.

Aproximadamente 75 tareas de desempeño y guías de calificación estarán disponibles a través del World Wide Web para ser utilizadas como herramientas instruccionales y de evaluación. Se exigirá que los maestros evalúen a todos sus alumnos cada año, utilizando tareas seleccionadas de esta base de datos. Se exigirá a todos los estudiantes del estado una tarea de desempeño común, para así calibrar las calificaciones que dan los maestros al trabajo del estudiante con los estándares estatales. Equipos de educadores de cada distrito calificarán los portafolios de los estudiantes cada otoño; los resultados serán tomados en cuenta en la instrucción en el aula. Examinadores estatales volverán a calificar una muestra de los portafolios de cada distrito para asegurar consistencia en la calificación.

La Evaluación del Próximo Siglo forma parte de un esfuerzo mayor por la responsabilización en Carolina del Norte que incluye el establecimiento de estándares altos; la creación de un sistema para basar las decisiones sobre promoción, retención y graduación en el desempeño estudiantil (acabando así con la promoción social o automática); y requisitos modificados para la graduación. La Comisión para los Estándares Educativos y la Responsabilización del Estado está desarrollando ahora un plan para introducir sus recomendaciones a lo largo de los próximos cuatro años, incluyendo lineamientos para el desarrollo profesional y la formación docente. A la fecha, el Consejo Estatal de Educación ha ordenado a la Superintendencia de Instrucción Pública que identifique los grados que servirán como puntos de corte en que los estudiantes deberán haber cumplido los estándares estatales para poder ascender al siguiente grado o nivel de estudio. También están en curso en la primavera de 1997 las pruebas de campo para el examen integral de secundaria y el examen de conocimientos esenciales.

Al mismo tiempo, muchos directores y profesores reportan preocupaciones serias sobre las demandas que las nuevas evaluaciones les imponen a ellos mismos y a sus escuelas (Aschbacher, 1993). Particularmente, informan que los maestros necesitan tiempo para familiarizarse con los nuevos estándares, evaluaciones y requerimientos administrativos; para comprender cómo se desarrollan y se califican las nuevas formas de evaluación; para aplicar los criterios en la evaluación del trabajo de sus estudiantes; y para adquirir suficiente información y conocimientos pedagógicos como para cambiar sus prácticas.

Efectuar cambios significativos en la práctica docente no es fácil ni barato. En un análisis sobre dos estados hecho por Lorraine Mc Donell, los maestros no entendían por completo las demandas de los estándares estatales y fueron incapaces de identificar actividades de aula correctamente alineadas, a pesar que se les ofreció desarrollo profesional y entrenamiento (McDonell & Choisser, 1997).

¿Quién debe ser responsable del desarrollo profesional en estos casos? ¿Quién debe pagar por ello? Al permanecer callados ante estas interrogantes, muchos estados han empujado la responsabilidad al nivel local. El supuesto, en este caso, parece ser que las estructuras de incentivos y responsabilización incitarán a los distritos escolares a proporcionar el apoyo adecuado.

Este supuesto puede ser razonable para aquellos distritos con recursos y capacidades necesarias. Sin embargo, en muchos distritos no existe apoyo para materiales nuevos o para el desarrollo profesional. Según el estudio de Mary Lee Smith sobre el hoy extinto Programa de Evaluación del Estado de Arizona (ASAP), el progreso más dramático ocurrió en escuelas que ya estaban cambiando, y lo más probable es que hubieran cambiado de cualquier manera (Smith, 1996). Las escuelas que no tenían el deseo o la capacidad de cambiar no se beneficiaron del programa ASAP. (Una de estas escuelas era geográficamente remota y contaba con pocos recursos; otra consideraba que sus estudiantes eran incapaces de alcanzar objetivos más altos.) (Ver McLaughlin, 1987; McDonell & Choisser, en prensa). En última instancia, los mandatos estatales no ofrecen panacea alguna.

12. La relación con las evaluaciones distritales y con las del aula

Una evaluación que intenta desempeñar demasiadas funciones -- diagnóstico del estudiante, planeamiento curricular, evaluación de programas, mejoramiento instruccional, responsabilización, certificación, comunicación pública -- inevitablemente, no hará ninguna bien. De allí que es importante distinguir roles apropiados para diferentes evaluaciones a nivel de distrito, escuela y salón de clases.

Al mismo tiempo, estas evaluaciones deben estar alineadas entre sí y con los estándares a los que se ciñen. Un sistema consistente asegura que los maestros y estudiantes comprendan qué es importante aprender y qué tan bien lo están haciendo.

La Ley de Educación para el Siglo Veintiuno de Oregon es un buen ejemplo. Según la ley, los estudiantes de todo el estado en los grados 3, 5, 8 y 10 deben rendir una serie de pruebas y evaluaciones locales uniformes, basadas en los estándares de contenido y desempeño establecidos en cada uno de estos grados. Las pruebas estatales incluyen preguntas de opción múltiple, ensayos y solución de problemas matemáticos. Las locales incluyen asignaciones en clase y otras tareas menos fáciles de evaluar. Mientras que estas tareas varían de profesor en profesor y de colegio en colegio, todos los estudiantes deben completar cierto número de tipos específicos de evaluaciones y obtener un puntaje mínimo.

Oregon ha establecido una guía de calificación de seis puntos para cada área de contenido, que debe ser utilizada al evaluar el trabajo de los estudiantes. El estado exige que los estudiantes logren alcanzar los estándares del décimo grado (en inglés, matemáticas, ciencias, historia, entre otros) para recibir un Certificado de Dominio Inicial. Aquéllos que cumplen los estándares del grado doce reciben un Certificado de Dominio Avanzado⁶.

El programa de evaluación del estado de Nuevo México proporciona otro ejemplo. El estado exige las mismas pruebas normativas para todos los estudiantes del tercer, quinto y octavo grado; portafolios de redacción de los estudiantes de cuarto, sexto y (opcionalmente) octavo grado; un examen de competencia de secundaria; y una evaluación de lectura diseñada por el distrito para el primer y segundo grado, cuyos resultados deben ser reportados al estado.

Como último ejemplo, Utah administra una prueba normativa estandarizada para todos los estudiantes del quinto, octavo y onceavo grados. El estado también ofrece a los distritos un conjunto de pruebas de desempeño referidas a criterios para evaluar el logro del alumno, basado en el marco estatal⁷.

Como muestran estos ejemplos, los programas evaluativos estatales y locales pueden alinearse de diversas maneras. La inclusión que hace Oregon del trabajo en el aula, así como la evaluación de portafolios utilizada en varios otros estados⁸,

⁶ Esta descripción se refiere a los requerimientos para el certificado de dominio inicial, ya que los concernientes al certificado de dominio avanzado están aún en elaboración. Ver *Metas curriculares comunes adoptadas: Estándares de contenido y desempeño y guías de calificación*. Departamento de Educación de Oregon, octubre de 1996.

⁷ Para ejemplos adicionales y mayor detalle, ver Bond, L.A., Braskamp, D., y Roeber, E. *Base de Datos de los Programas estatales de evaluación estudiantil del año escolar 1994-1995*. Oakbrook, IL: Laboratorio Educativo de la Región Norcentral / Consejo de Funcionarios Principales de las Escuelas Públicas, 1995.

⁸ Un portafolio es una colección de trabajos de un estudiante diseñado para mostrar su progreso a lo largo del tiempo y su nivel de logros.

representan intentos explícitos por acoplar los estándares y evaluaciones estatales a las prácticas en el aula. Este acoplamiento es crítico para el logro del estudiante.

La evaluación, por supuesto, es una parte integral del proceso de enseñanza-aprendizaje, que ocurre continuamente en las prácticas del aula. Los maestros usan de manera rutinaria una amplia variedad de evaluaciones formales (exámenes, pruebas, tareas para la casa, monografías semestrales, proyectos), así como medios más informales (preguntas orales, discusiones en clase, observación de las expresiones faciales de los estudiantes) para medir el progreso del alumno, asignar calificaciones, motivar la atención, proporcionar retroalimentación y adaptar la instrucción a las necesidades de los alumnos.

De manera similar, los estudiantes se dedican regularmente a auto-evaluaciones informales mientras estudian e intentan resolver problemas, monitorean su propio progreso y mejoran su aprendizaje. De hecho, tanto maestros como estudiantes pasan mucho más tiempo dedicados a la auto-evaluación que a completar pruebas externas. La auto-evaluación también ejerce mayor influencia en las decisiones instruccionales diarias de los maestros y en las experiencias de aprendizaje de los alumnos. Tanto las prácticas en el aula como la auto-evaluación proporcionan la comprensión detallada y la retroalimentación continua que profesores y estudiantes necesitan para guiar un aprendizaje efectivo y continuo.

Es importante, debido a su rol fundamental, que la práctica de la evaluación en clase y las auto-evaluaciones de los estudiantes sean guiadas por los mismos estándares en los cuales están basadas las otras evaluaciones. Las evaluaciones externas pueden ser de ayuda en este sentido, porque sirven como modelos y porque ayudan a los profesores a comprender los nuevos estándares de desempeño de los estudiantes. (Las guías de calificación de Oregon representan justamente estas herramientas). Es igualmente importante el apoyo continuo al desarrollo profesional.

Conclusión

“Una imagen clara de lo que se desea lograr, un sistema de medición integral para evaluar el progreso, y un compromiso de actuar según los resultados para realizar los cambios apropiados”. Esos fueron los requisitos formulados por el gobernador Romer para el mejoramiento del desempeño de los estudiantes⁹. Se pretende que los estándares de contenido y de desempeño proporcionen la “imagen clara” de lo que se necesita lograr. Las evaluaciones sólidas alineadas con esos estándares constituyen el “sistema de medición.” Demostrar el “compromiso de actuar”, proporcionando recursos instruccionales de alta calidad y desarrollo profesional amplio, comprometiendo a todos los alumnos y asegurando un amplio apoyo y compromiso públicos, ese es el reto que afrontamos.

⁹ Declaración del Gobernador Romer citada por el Panel de Metas Educativas de Colorado. *Alianzas para educar a los estudiantes de Colorado: Sacando lo mejor de todos nuestros estudiantes*. 1995.

BIBLIOGRAFIA

- American Federation of Teachers (AFT). (1996). *Making standards matter* (p.19). Washington D.C: AFT.
- Aschbacher, P.R. (1993). *Issues in innovative assessment for classroom practice: Barriers and facilitators* (Reporte Técnico de la CSE No. 359). Los Angeles: Centro Nacional para la Investigación en Evaluación, Estándares y Medición del Logro Estudiantil de UCLA (CRESST).
- Baker, E.L. (1994). *Researchers and assessment policy development: A cautionary tale*. **American Journal of Education**. 102(4), 450-478.
- Bond, L.A., Braskamp, D., & Roeber, E.(1996). *The status report of the assessment programs in the United States*. Oakbrook, IL: NCREL / Council of Chief State School Officers.
- Corbett, H.D., & Wilson, B.L.(1991). *The central office role in instructional improvement*. Philadelphia, PA: Investigación para Mejores Escuelas, Inc. (Servicio de Reproducción Documental ERIC No. ED 374567)
- Cremin, L.A. (1989). *Popular education and its discontents*. (pp.9). Nueva York: Harper y Row.
- Diegmuller, K. (2 de noviembre de 1994). "Panel unveils standards for history: Release comes amid outcries of imbalance" en **Education Week** 14 (9), 1, 10.
- Dorr-Bremme, D.W., & Herman, J.L. (1986). **Assessing student achievement: a profile of classroom practices**. Serie de monografías sobre Evaluación del CSE No. 11. Universidad de California, Los Angeles, Centro para el Estudio de la Evaluación. (Servicio de Reproducción de Documentos ERIC No. ED 338691)
- Dunbar, S.B., Koretz, D.M., & Hoover, H.D. (1991). *Quality Control in the development and use of performance assessments* en **Applied Measurement in Education** 4, (4), 289-303.
- Herman, J.L. y Golan, S. (1991). *Effects of standardized testing on teachers and learning - Another look* (Reporte Técnico del CSE No.334). Los Angeles: Centro Nacional para la Investigación en Evaluación, Estándares y Medición del Rendimiento Estudiantil de UCLA (CRESST).
- Herman, J.L., Aschbacher, P.R. y Winters, L.(1992). *A practical guide to alternative assessment*. Alexandria, VA: Asociación de Supervisión y Desarrollo Curricular.
- Kellaghan, T. y Madaus, G.F. (1991). "National testing: Lessons for America from Europe. (The testing issue) en **Educational Leadership**. 49(3), 87-94.
- Kentucky Department of Education (KDE). (1994). **Kentucky Instructional Results Information System, Reporte Técnico de 1992-93**. Frankfort, KY: KDE.
- Koretz, D.M., Baron, S., Mitchell, K.J. y Stetcher, B.M. (1996). *Perceived effects of the Kentucky Instructional Results Information System (KMIS)*. Santa Mónica, CA: RAND.
- Koretz, D.M., Mitchell, K.J., Baron, S. Y Keith, S. (1996). *Perceived effects of the Maryland State Assessment Program*. (Reporte Técnico del CSE No. 409). Los Angeles: Centro para el Estudio de la Evaluación de UCLA.

Koretz, D.M., Stetcher, B., Klein, S., McCaffrey, D., & Deibert, E. (1993). *Can portfolios assess student performance and influence instruction? The 1991-92 Vermont Experience* (Reporte Técnico del CSE No. 371). Los Angeles: Centro Nacional para la Investigación en Evaluación, Estándares y Medición de Logros Estudiantiles de UCLA (CRESST).

Linn, R.L., Burton, E., De Stefano, L. y Hanson, M. (1995). *Generalizability of New Standards Project 1993 pilot study tasks in mathematics* (Reporte Técnico del CSE No. 392). Los Angeles: Centro Nacional para Evaluación, Estándares y Medición del Rendimiento del Alumno de UCLA (CRESST).

Linn, R.L. y Baker, E.L. (1996). "Can performance-based assessments be psychometrically sound?" en J.B. Baron y D.P. Wolf (Eds.), ***Evaluación basada en el desempeño del estudiante: Desafíos y Posibilidades. Anuario 87 de la Sociedad Nacional para el Estudio de la Educación.*** Parte I (pp. 84- 103). Chicago: Editorial de la Universidad de Chicago.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). "Complex performance-based assessment: expectations and validation criteria" en ***Educational Researcher*** 20, 15-21.

Madaus, G.f. (1988). "The influence of testing on the curriculum" en L. N. Tanner (ed.); ***Problemas críticos del currículo, 87avo. Anuario de la Sociedad Nacional para el Estudio de la Educación.*** Parte I (pp. 83-121). Chicago: Editorial de la Universidad de Chicago.

McDonell, L.M. (1994). *Policymaker's views of student assessment* (Reporte Técnico No. 378). Los Angeles: Centro Nacional para la Evaluación, Estándares y Medición del Rendimiento Estudiantil de UCLA (CRESST).

McDonell, L.M. (1996). *The politics of state testing: Implementing new student assessments* (Reporte Técnico) (p.31). Los Angeles: Centro Nacional para la Evaluación, Estándares y Medición del Rendimiento Estudiantil de UCLA (CRESST).

McDonell, L.M. y Choisser, C. (en prensa). *Testing and teaching: Local implementation of new state assessments.* Los Angeles: Centro para el Estudio de la Evaluación de UCLA.

McLaughlin, M.W. (1987). *Learning from experience: lessons from policy implementation* en ***Educational Evaluation and Policy Analysis.*** 9(2), 171-178.

National Academy of Education (NCE). (1996). *Quality and utility of the 1994 Trial State Assessment in Reading* (chapter 4). Stanford, CA: Universidad de Stanford, NCE.

National Council of Teachers of Mathematics (NCTM). (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: National Council of Teachers of Mathematics.

National Council of Teachers of Mathematics. (NCTM). (1995). *Assessment standards for school mathematics.* Reston, VA: National Council of Teachers of Mathematics.

National Education Goals Panel. (1993). *Report of goals 3 and 4, technical planning group on the Review of Education Standards*. Washington, DC: National Education Goals Panel.

Phillips, S.E. (1995). Legal defensibility of standards: Issues and policy perspectives. *Actas de la conferencia conjunta sobre el establecimiento de estándares para evaluaciones a gran escala de la Junta Directiva para Pruebas Nacionales (NAGB) y del Centro Nacional de Estadísticas Educativas (NCES), Vol. II* (pp.379-393). Washington, DC: NAGB y NCES.

Resnick, L.B., & Resnick, D.P. (1992). "Evaluating the thinking curriculum: new tools for educational reform" en B.R. Gifford y M.C. O'Connor (Eds.), **Changing assessments: Alternative views of aptitude, achievement and instruction**. (pp. 37-75). Boston: Editores Académicos Kluwer.

Shavelson, R.J., Gao, X., & Baxter, G.P. (1993). *Sampling variability of performance assessments*. Reporte Técnico de la CSE No. 361. Los Angeles: Centro Nacional para la Evaluación, Estándares y Medición del Rendimiento Estudiantil de UCLA (CRESST).

Shavelson, R.J., Baxter, G.P. y Pine, J. (1991) *Performance assessments: Politics of achievement measurement*. Ponencia invitada a la conferencia sobre Mehrdimensionale Lehr- Lern- Arrangements: Lernen, Denken, Handeln in Komplexen Okomischen Situationen, Gottingen, Alemania.

Shavelson, R.J., Mayberry, P.W. y Webb, N. (1990). *Generalizability of job performance measurements: navy machinists mates* en **Military Psychology** 2, 129-144.

Shepard, L.A. (1991). "Will national tests improve student learning?" en **Phi Delta Kappan** 73 (3), 232-239.

Smith, M.L., Edelsky, C., Draper, K., Rottenberg, C. y Cherland, M. (1991). *The role of testing in elementary schools*. Reporte Técnico No.321 de la CSE). Los Angeles: Centro Nacional para la Evaluación, Estándares y Medición del Rendimiento Estudiantil de UCLA(CRESST).

Smith, M.L. (1996). Reforming schools by reforming assessment: Consequences of the Arizona Student Assessment Program. Reporte Técnico del CSE. Los Angeles: Centro Nacional para La Evaluación, Estándares y Medición del Rendimiento Estudiantil de UCLA (CRESST).

Webb, N. (1997). *Determining alignment of expectations and assessments in mathematics and science education* en **Resúmenes del Instituto Nacional para la Educación Científica (NISE)**, 1 (2), enero.

Wiley, D. (en prensa). The New Standards Reference Examination Standards-Referenced Scoring System. Los Angeles: Centro para el Estudio de la Evaluación de UCLA.

Wolf, S.A. y Gearhart, M. (1993). **Writing what you read: Assessment as a learning event**. (Reporte Técnico No. 358). Los Angeles: Centro Nacional para la Investigación en Evaluación Estándares y Medición del Rendimiento Estudiantil de UCLA (CRESST)

Centro Nacional para la Investigación sobre Evaluación, Estándares y Medición del Rendimiento Estudiantil (CRESST)
Centro para el Estudio de la Evaluación (CSE)
Escuela de Graduados en Estudios de Educación e Información
Universidad de California, Los Angeles, CA 90024 – 6511
(310) 206 – 1532
Copyright © 1997 The Regents of the University of California.
Febrero de 1997

Este trabajo estuvo apoyado por un donativo a la Comisión de Educación de los Estados realizado por la Fundación Nacional de la Ciencia (*National Science Foundation, NSF*), número REC-915439, Jane Armstrong, investigadora principal. Las opiniones, conclusiones o recomendaciones expresadas en este material son de los autores y no reflejan necesariamente el punto de vista de la NSF. El trabajo reportado en esta publicación también estuvo apoyado por el Programa de Centros de Investigación y Desarrollo Educativo, donación número R305B600002, administrada por la Oficina de Investigación y Mejoramiento Educativo del Departamento de Educación de los Estados Unidos. Los hallazgos y opiniones que aquí se expresan no reflejan la posición o las políticas del Instituto Nacional de Logros Estudiantiles, la Oficina de Investigación y Mejoramiento Educativo o el Departamento de Educación de los Estados Unidos.

Otra versión de este manuscrito, más orientada a políticas, titulada “Guía sobre la evaluación impulsada por estándares para formuladores de políticas” puede obtenerse de la Comisión de Educación de los Estados (*Education Commission of the States - ECS*), Centro de Distribución ECS, 707 17th Street, Suite 2700, Denver, CO 80202 – 3427, 303 – 299 – 3692. pedir N. SI –97 – 3.