Pruebas y Rendición de Cuentas

Felipe Martínez Rizo

CUADERNO No. 12

Instituto Nacional para la Evaluación de la Educación

Colección Cuadernos de Investigación

ISSN 1665-9457

Pruebas y Rendición de Cuentas

Felipe Martínez Rizo*

CUADERNO No. 12

*Director General del Instituto Nacional para la Evaluación de la Educación

Este es un resumen ejecutivo de su estudio, el texto completo puede consultarse en: www.inee.edu.mx

MÉXICO, JULIO, 2003

Contenido

Introducción

La evaluación es una actividad que ocupa un lugar importante en el trabajo educativo. La valoración del avance de los alumnos en cada materia o área del currículo y grado, debe ocurrir permanentemente, si se quiere que discípulos y docentes tengan la retroalimentación necesaria para ajustar el proceso de aprendizaje y enseñanza. Los padres de familia necesitan también información clara sobre el avance de sus hijos y sobre el funcionamiento de la escuela.

Pero las evaluaciones que hacen los maestros, aunque indispensables para que haya buena enseñanza, tienen limitaciones: no son comparables entre sí, ni siquiera en una misma escuela, mucho menos a escala regional, nacional o internacional.

El mundo actual, por su parte, con sus tendencias globalizadoras, plantea exigencias nuevas a las naciones en lo político, social, económico, científico y tecnológico, industrial, comercial y financiero.

La globalización trajo consigo, entre otros aspectos, la necesidad de contar con evaluaciones comparables en el plano internacional. Así comenzaron a desarrollarse, desde los años sesenta del siglo XX, proyectos internacionales de evaluación educativa, basados en las metodologías previamente desarrolladas en los sistemas educativos de algunos países, en especial los Estados Unidos, como se mencionará más adelante.

Más reciente ha sido el surgimiento de un gran interés, en amplios sectores de la sociedad, por la calidad educativa y, más precisamente, por la evaluación de dicha calidad. Muestra de ese interés es el impacto en los medios de comunicación de muchos países, de los resultados de evaluaciones educativas internacionales, como las pruebas PISA de la OCDE, o las pruebas de los Estudios Internacionales de Matemáticas y Ciencias (TIMSS), cuyas versiones sucesivas desde los años sesenta habían pasado inadvertidas en nuestro medio.

Pero la evaluación educativa comprende una amplia gama de actividades, según los *sujetos* (que pueden ser alumnos, docentes, currículo, escuelas, sistema educativo como tal, entre otros), los *aspectos* a evaluar, los *propósitos* de la evaluación, etcétera. Si se tienen en cuenta, además, las dimensiones del sistema educativo en un país como México, y la vaste-

dad de los contenidos del currículo de la educación básica, se entiende que la evaluación educativa es una tarea sumamente compleja también, por lo que debe ser realizada por diversas instancias, en forma articulada.

La evaluación demanda mecanismos específicos para los diversos tipos educativos (superior, media superior y básica); comprende pruebas para valorar conocimientos, aptitudes y actitudes, así como indicadores y evaluaciones comprensivas; incluye evaluaciones individuales y pruebas de aplicación masiva; evaluaciones de alumnos, maestros, directivos, escuelas, zonas y sistemas estatales; entre otras dimensiones a considerar.

De especial interés, por la importancia que están adquiriendo en muchos países, es lo relativo a las evaluaciones que se hacen en gran escala. La creciente atención que prestan a estas formas de evaluación educativa los responsables de las políticas públicas, medios de comunicación y sociedad en general, ha hecho surgir demandas específicas para que los resultados se difundan amplia y oportunamente, poniéndose al alcance no sólo de autoridades educativas, sino también de maestros, padres de familia y toda persona interesada en ellos.

En principio, estas demandas son lógicas y legítimas, pero no lo son con igual claridad en todos los detalles. Parece haber consenso en dos puntos: por una parte, en que deben difundirse los resultados agregados a nivel nacional y estatal; por otra, en que no deben darse resultados de alumnos individuales.

El último punto puede parecer extraño, pero no lo es si se profundiza en la materia: es clara la necesidad de cada alumno de conocer los resultados de sus propias evaluaciones, y obvio también el derecho de los padres de ser informados sobre los resultados obtenidos por sus hijos. Una mínima comprensión del derecho a la privacidad bastará, sin embargo, para entender que los resultados individuales no deben ser del dominio público.

En cambio, no es clara la conveniencia de dar resultados de cada escuela en lo individual. En México, una parte de las demandas sociales recientes más visibles en los medios de comunicación se refiere a este punto, al parecer sin una comprensión suficiente de su complejidad.

Pros y contras de la difusión de resultados por escuela

A favor de la difusión de resultados por escuela se aduce el argumento de que, para propósitos de mejora, es insuficiente dar resultados a nivel nacional o estatal, los cuales son demasiado generales, con promedios que comprenden diferencias considerables de las escuelas que forman parte del sistema o subsistema del cual se trate.

En relación con los padres de familia que tienen a sus hijos en escuelas privadas, se señala que conocer los resultados por escuela es necesario para sustentar la decisión de cuál escuela seleccionar para enviar a sus hijos; en el caso de quienes tienen a sus hijos en

escuelas públicas, en especial cuando no hay posibilidades reales de elección, se señala también que sólo conociendo los resultados de la escuela en que están sus hijos podrán los padres de familia exigir de las autoridades el mejoramiento de la calidad y corrección de las deficiencias existentes.

Sin negar la fuerza de los argumentos anteriores, desde otro punto de vista se subrayan los riesgos que puede tener en algunos casos la difusión de resultados de evaluación por escuela. Se aducen en particular dos razones:

- En evaluaciones basadas en muestras, que no cubren a todas las escuelas de un sistema, sino sólo una fracción relativamente pequeña de ellas, los resultados pueden sustentar juicios consistentes sobre el conjunto del sistema si la muestra es adecuada; pero las escuelas que obtengan los resultados más altos o más bajos no pueden considerarse por ello, obviamente, *las mejores o peores* del sistema.
- Por otra parte, aún en el caso de evaluaciones censales, se advierte sobre lo inadecuado de ordenar a las escuelas en una clasificación que las sitúa como buenas o malas (ranking), con base en los resultados obtenidos por sus alumnos en las pruebas de aprendizaje, aplicadas al final del ciclo escolar más reciente.

Dadas las limitaciones técnicas de cualquier prueba, una evaluación de escuelas en lo individual basada en el rendimiento de los alumnos tendría que basarse en varias evaluaciones, para evitar el riesgo de juzgar con base en un solo resultado, el cual podría no ser representativo de la situación real de cada plantel.

La motivación de los alumnos para responder una prueba, por ejemplo, puede variar de una aplicación a otra. Las fluctuaciones derivadas de cambios en diversas circunstancias de las aplicaciones de pruebas, pueden producir cambios en los resultados que no tienen que ver con la competencia real de los alumnos; eso es lo que se conoce como *volatilidad* de los resultados.

Importa también controlar el peso de los factores del entorno, y enriquecer la perspectiva de las pruebas con valoraciones de los procesos que tienen lugar en cada escuela, para eso se requieren acercamientos de tipo cualitativo.

En vez de resultados que sólo reflejen el aprendizaje alcanzado por los alumnos en el momento de aplicación de una prueba, convendrá tener evaluaciones longitudinales, que permitan valorar el avance de los alumnos en el logro de los objetivos educativos, la llamada *ganancia*.

Controlando los factores del contexto y con mediciones de ganancia, podrá valorarse también el *valor agregado* por una escuela en lo que se refiere al aprendizaje de los alumnos, a diferencia de los avances no atribuibles a ella.

Sin esas condiciones, la difusión de resultados por escuela, sobre todo en forma de ordenamientos simples, puede ser negativa e injusta para escuelas cuyo alumnado está formado mayoritariamente por alumnos de medios desfavorecidos.

Por la importancia que tiene el tema para las políticas educativas de México, en los apartados siguientes se presentan elementos adicionales de juicio que permitan sustentar una posición más adecuada al respecto.

Pruebas, difusión de resultados y tablas de posiciones: algunas experiencias

Las tendencias prevalecientes en muchos países en la actualidad, incluyen demandas sociales en dirección de una rendición de cuentas cada vez más clara, lo cual permita a la sociedad apreciar la manera en que se emplean los recursos derivados de los impuestos en los servicios públicos.

En algunos sistemas educativos, como los del Reino Unido, Francia y Chile, o algunos estados de Australia y la Unión Americana, esas demandas han llevado al establecimiento de sistemas de rendición de cuentas (*accountability*), basados en los resultados de pruebas de rendimiento aplicadas a los alumnos: se presentan los puntajes promedio obtenidos por los alumnos de cada escuela, en forma similar a las tablas de posiciones de los equipos de una liga deportiva (*rankings, league tables*).

Una visión optimista pero simple de esos sistemas, los ve como una solución casi mágica de todos los problemas de la educación. La experiencia real de los lugares en donde se han aplicado exige mayor cautela. Incluso en donde no se publican puntajes crudos, sino medidas más refinadas de valor agregado, el uso de *rankings* hace inevitable que haya ganadores y perdedores, y una vez que una escuela es etiquetada como deficiente, es difícil encontrar la manera de ayudarla en un ambiente donde prevalece la postura de vergüenza y recriminación, que no lleva a la formulación de estrategias de mejora.

En América Latina, Chile destaca como el país que ha desarrollado el sistema de evaluación educativa más completo, como parte de un esfuerzo mayor de reforma educativa, desde fines de los años ochenta. Los resultados recientes de las evaluaciones de dicho sistema (*Sistema de Medición de la Calidad Educativa*, Simce), sin embargo, mostraban un estancamiento de los niveles educativos y no la mejora esperada.

Los resultados de las pruebas PISA Plus, difundidos en julio de 2003, mostraron también que los niveles de aprendizaje de los jóvenes chilenos de 15 años, se situaban por debajo tanto de los argentinos como de los mexicanos. No es casual, por ende, que el gobierno nombrara una comisión para estudiar el tema y presentar recomendaciones sobre el uso y el desarrollo del Simce, (Cfr. Schiefelbein, 2003).

Reflexionando sobre lo anterior, estudiosos chilenos hicieron interesantes consideraciones sobre el papel de las evaluaciones y la complejidad de los procesos de mejora educativa. Comentando las reacciones de la prensa y los medios políticos ante los resultados del Simce, Pablo González, de la Universidad de Chile, señala:

La reciente publicación de los resultados del Simce*... ha generado un interesante debate en la prensa que refleja la importancia que se le asigna al tema. En forma predecible, algunas de estas opiniones son críticas y tienen un carácter político... La sola posibilidad de emitir estos juicios es muy valiosa, y es importante reconocer la actitud del ministerio de hacer públicas las mediciones (en algunos países ni siquiera pueden ser conocidas por los propios profesores de cada escuela), mejorar su comparabilidad (para poder medir cambios en el tiempo) y difundirlas lo más rápidamente posible. También es valiosa la participación en varias mediciones comparables internacionalmente...*

Frente al camino fácil de denostar a todos los actores del sistema escolar por los bajos resultados, es necesario aclarar que el objetivo de participar en las mediciones internacionales no era ganar medallas sino saber cómo estábamos. Si se sabía que, dadas las naciones participantes, Chile iba a ranquear en la parte más baja del ranking, ¿era la decisión políticamente más sensata no ser medidos? La respuesta a esta pregunta depende de la madurez de los países. Quizás es la calidad de la política y de la prensa la que se está midiendo, a través de las reacciones que suscitan la divulgación de las distintas mediciones. El objetivo de participar en el TIMSS*, en* CIVIC *o en el* IALS *no era deprimirnos, sino aprender.*

En educación se recurre a las pruebas tipo Simce *principalmente porque la calidad de la enseñanza que ofrece una escuela o un país no es directamente observable. Las familias y los países pueden ganar al contar con esta información. Lamentablemente también puede ocurrir exactamente lo contrario.*

Una primera forma de que ocurra esto último es rechazar todo lo que se ha venido haciendo en los últimos años y comenzar a proponer (y peor aún ensayar) nuevos caminos en forma apresurada...Otra forma de que los países y las familias pierdan con la publicación del Simce *es que las escuelas decidan mejorar sus resultados en las evaluaciones excluyendo a los alumnos con mayores problemas.*

Esto de hecho lo están haciendo muchos colegios, y es muy importante que los sistemas de evaluación se hagan cargo de estos incentivos incorrectos que generan. Obviamente la solución a este problema no es dejar de evaluar, ya que ello hace imposible superar el problema de baja calidad del que hemos tomado masivamente conciencia.

...Una propuesta que se ha considerado para reducir los incentivos a que las escuelas segreguen es medir "valor agregado", esto es evaluar al mismo alumno al comienzo y al final de un ciclo de enseñanza, haciendo público el mejoramiento y no el resultado bruto. Se menciona la experiencia neozelandesa, donde se evalúa a los alumnos al ingresar en primero básico y luego se mide su progreso a través de cada ciclo (no cada año). Otra medida complementaria ha sido adoptada por Suecia, al vincular el financiamiento de los establecimientos con el egreso oportuno de sus alumnos de la secundaria.

...Otra lección en el mismo sentido que arrojan los estudios sobre funciones de producción, y esto se ha reiterado en la discusión en torno al Simce*, es que el nivel socioeconó-*

mico de las familias es un determinante fundamental del puntaje que pueden alcanzar los alumnos en este tipo de pruebas. Si nos interesa construir una sociedad con una mayor igualdad de oportunidades, es necesario hacerse cargo de esta reproducción social de las desigualdades e incrementar el valor de la subvención pagada por los estudiantes de hogares más pobres.

Para alcanzar un mismo nivel de resultados y para que estos alumnos sean atractivos para las escuelas este paso es necesario. Si bien sus consecuencias sobre la equidad y la integración son evidentes, lo que no está claro es cuál será su efecto sobre la eficiencia del sistema. Debido a este posible trade-off eficiencia y equidad y a la escasez de recursos públicos, lo más indicado en esta materia es, como se ha sugerido respecto a todos los programas, experimentar y asegurar las correcciones que aseguren que los mayores recursos rindan frutos.

Volvemos así a nuestro tema inicial. ¿Se puede decir que ha sido realmente ineficiente el mayor gasto en educación por la falta de progreso en el Simce*? Habría que responder primero algunas interrogantes. No sabemos cuánto demora este tipo de esfuerzo en mostrar resultados... países como Estados Unidos, Francia y Japón han elevado en forma sustancial sus presupuestos en educación; sin embargo, sus resultados se mantienen prácticamente estancados no sólo en períodos breves sino a lo largo de varias décadas.*

...En realidad la pregunta no es a quién echarle la culpa, sino cuál es la responsabilidad que cada uno puede asumir para cambiar la situación. Tengamos una perspectiva más amplia. Entendamos que los cambios culturales son los que toman más tiempo. Y que el tema principal no es un problema de recursos... otros países que gastan lo mismo que nosotros se ubican entre los países con mejores rendimientos.

Debemos acercarnos a las verdaderas claves, evaluar lo que funciona bien, aprender y replicar. Si el gobierno no hubiese centrado (correctamente y con gran responsabilidad política) la discusión en la calidad de la educación quizás ahora estaríamos festejando los indicadores más tradicionales de cobertura, deserción, repitencia y años de escolaridad de la población, que muestran mejoramientos sustanciales... No es verdad que la situación fue mejor hace veinte o treinta años, cuando apenas una fracción de la población terminaba la enseñanza media. Pero podemos obtener un sistema educacional como soñamos –que conjugue calidad, equidad y diversidad–, en una o dos décadas, si, unidos por esa visión, somos suficientemente pacientes, perseverantes y capaces de aprender y corregir, (González, 2003).

En el marco de este debate chileno, un experto de gran prestigio internacional, el exministro de educación de ese país Ernesto Schiefelbein, enriquece la reflexión sobre los propósitos de los sistemas de evaluación en gran escala y sobre la diferencia de uno que busque propiciar el mejoramiento, frente a los que se limitan a la rendición de cuentas:

Desde fines de los año sesenta se ha medido el nivel de aprendizaje de los alumnos para elevarlo, pero no ha sido evidente la forma de hacerlo y, de hecho, no ha aumentado. La prueba de 8° grado (1967-1971) se limitó a mostrar a los profesores los tipos de habilidades que debían estimular en sus alumnos. La PER*, creada a principios de los años ochenta, buscó determinar el nivel del servicio educativo que se ofrecía y "monitorearlo" con ayuda de los padres, que tratarían de seleccionar las mejores escuelas para sus hijos. El* Simce *midió, a partir de 1988, los rendimientos de las escuelas y trató de identificar factores que pudieran explicar las diferencias y evaluar el impacto de los programas de los municipios y del* Mineduc*. Sin embargo, no hay avances y conviene repensar este poderoso instrumento, que es el* Simce*, si se quiere mejorar los aprendizajes. Un rol alternativo del* Simce *sería, según otros expertos, ayudar a los profesores a enseñar mejor y revisar cuáles son los temas importantes del currículo, más que el* clasificar *rendimientos. Este segundo rol plantea diferencias nítidas con respecto al primero: ¿Seleccionar o desarrollar? ¿Poner nota o medir logro de meta? ¿Conocer el ranking o el aprendizaje de objetivos específicos (criterio)? ¿Castigar o ayudar?*

Son dos funciones diferentes (los expertos hablan de Evaluación Sumativa y Formativa, respectivamente). Si la primera no elevó hasta ahora los rendimientos, convendría probar la segunda. Esto implica fuertes cambios en el Simce*. En efecto, para* seleccionar *o* castigar *basta calcular un puntaje total que* discrimine entre buenos y malos *y difundirlo públicamente. En este rol es fundamental no divulgar los ítemes utilizados para que mantengan su poder de discriminación (no ser conocidos por los profesores y evitar que preparen a sus alumnos para contestarlos mecánicamente). Basta calcular el puntaje de cada escuela para que los padres seleccionen entre escuelas o preparar datos e indicadores globales para que los funcionarios del Mineduc modifiquen las estrategias.*

En la segunda función, en cambio, el usuario de la información del Simce *es el profesor en su sala de clases. En efecto, para* desarrollar *o* ayudar a mejorar *se necesita entregar información detallada a cada profesor sobre los aspectos en que cada uno de sus alumnos logró los niveles adecuados y aquellas habilidades o conocimientos que todavía no posee o no domina suficientemente. El profesor debe saber lo que contestó cada alumno en la prueba y revisar con cada uno los errores que cometió, hasta que el alumno internalice las deficiencias y reforme adecuadamente sus procesos de pensamiento. Sin información detallada el profesor no puede identificar los aspectos de su enseñanza que debe cambiar o cómo ayudar a cada estudiante.*

Usar el Simce *para que cumpla esta segunda función implica cambios importantes. Por ejemplo, divulgar tanto el puntaje total en lenguaje como puntajes parciales en aspectos específicos, tales como: comprensión de lectura, completar frases, vocabulario, analogías, antónimos, ortografía o gramática. Son estos puntajes parciales los que permiten al profesor determinar qué aspectos privilegiar o modificar...*

Pero, además, habría que particularizar los datos que tendrían una difusión pública y los que se entregarían en forma privada. Por ejemplo, entregar en forma privada *a cada profesor los resultados de cada uno de sus alumnos, para que él reflexione sobre las estrategias específicas a usar con cada uno de ellos. Al mismo tiempo, convendría continuar difundiendo públicamente sólo los puntajes totales de cada escuela.*

Sería oportuno, además, que los especialistas que revisan periódicamente el currículo usaran esta información pormenorizada del Simce*. Sabrían, entonces, cuáles son aquellos aspectos que constituyen barreras para que el alumno continúe avanzando en sus aprendizajes. Analizar, además, el tiempo que toma dominar cada uno de los objetivos... Este análisis permitiría dar más tiempo para algunas materias o proporcionar recomendaciones o modelos (frameworks, scripts o guías) para que los profesores cuenten con elementos para reforzar sus estrategias de enseñanza en esas habilidades o conocimientos fundamentales que los alumnos no logran superar.*

En resumen, una redefinición clara del objetivo de la prueba nacional de medición de la calidad de la educación debe orientar la reflexión del grupo de expertos sobre los cambios que debe tener el Simce... (Schiefelbein, 2003)

La experiencia de los dos países mencionados, particularmente relevante en el campo que nos ocupa, apunta con claridad en el sentido de que un sistema de evaluación que se limite a clasificar u ordenar escuelas o alumnos, según su nivel de desempeño en ciertas pruebas, no contribuye por sí mismo a elevar la calidad educativa, como tampoco lo consiguen medidas coercitivas simples. Mejorar los niveles educativos es una tarea que requiere plazos largos y estrategias complejas. Para apoyarlas se necesitan sistemas de evaluación que, sin complacencia alguna, y gracias a una mejor comprensión de los procesos educativos, permitan juicios ponderados sobre el pasado y el presente, y sustenten expectativas razonables hacia el futuro, como punto de partida para el diseño y la implementación de estrategias de mejora efectivas.

El desarrollo histórico de los sistemas de evaluación en gran escala

El uso de pruebas de rendimiento académico de aplicación masiva es bastante antiguo, si no nos limitamos a los instrumentos estandarizados con las herramientas de la psicometría moderna. Desde mediados del siglo XIX comenzaron a usarse pruebas artesanales en algunos estados americanos, como Massachussets, con Horace Mann. La primera prueba estandarizada de rendimiento, el *Stanford Achievement Test*, fue aplicada en 1923.

Desde entonces hasta los años sesenta del siglo pasado, el uso de este tipo de pruebas se extendió casi exclusivamente en los Estados Unidos, y en ese país los resultados de las pruebas se empleaban solamente para decisiones relativas a alumnos en lo individual, y

no para evaluar sistemas educativos como tales. Además, salvo en el caso de las pruebas de ingreso a la educación superior, las pruebas no tenían consecuencias fuertes para el alumno.

A partir de finales de los años sesenta la situación comenzó a cambiar significativamente, con la creación del *National Assessment of Educational Progress* (NAEP), con las disposiciones que requerían que se evaluaran los programas compensatorios impulsados por el Título I de la Ley de Educación Elemental y Secundaria (ESEA), y con el movimiento de las pruebas de competencias mínimas de los años setenta, (Hamilton y Koretz, 2002).

Durante las dos últimas décadas del siglo XX el uso de pruebas de rendimiento estandarizadas en gran escala se extendió en muchos países del mundo, y en los Estados Unidos adoptó progresivamente nuevas características. Las pruebas se usan cada vez más para formular juicios evaluativos, tanto sobre alumnos individuales, como sobre escuelas singulares y sobre sistemas educativos en conjunto, en los niveles de distrito, estado o región y país; es notorio el uso creciente de las pruebas para comparaciones internacionales.

Las consecuencias de los resultados obtenidos en las pruebas tienden, además, a ser cada vez más fuertes: las decisiones sobre promoción o no promoción de cada alumno dejan de ser tomadas por cada maestro y cada escuela, para depender de los resultados en las pruebas estandarizadas; pero también las escuelas pueden ser premiadas o sancionadas con base en los resultados de sus alumnos, con la posibilidad incluso, de que se decida el cierre de un plantel con base en ello; en los países federales los sistemas de las entidades federativas pueden ser objeto también de medidas de apoyo o de presión a partir de los resultados de las pruebas; y en el nivel internacional, si bien no existen mecanismos de sanción con fuerza legal, está claro el impacto que tiene la difusión de los resultados de las evaluaciones en la opinión pública, y en las políticas gubernamentales.

La adopción de la última versión de la Ley de Educación Elemental y Secundaria, firmada por el Presidente Bush en enero de 2002, y conocida con la expresión, ya mencionada, de *No Child Left Behind* (NCLB), es la manifestación más clara y fuerte de la tendencia mencionada: en un plazo de cuatro años, todos los estados americanos deberán haber establecido un conjunto de estándares curriculares y un sistema de pruebas estatales alineadas a dichos estándares, que deberán aplicarse a todos los alumnos de los grados 3° a 8°.

Para asegurar la comparabilidad de los resultados, todos los estados deberán, además, participar en las evaluaciones nacionales del NAEP, lo que hasta ahora era opcional. La Ley NCLB establece, además, medidas de apoyo y sanciones, las cuales pueden llegar hasta la clausura, para las escuelas que cumplan o no las ambiciosas metas que se establecen, a las que ya se ha hecho alusión. Esta nueva forma de emplear los resultados de las evaluaciones es lo que trata de designar la expresión Sistemas de Rendición de Cuentas Basados en Pruebas (*Test-Based Accountability Systems*).

Los desarrollos que están teniendo lugar en México, en el campo de la evaluación educativa, se sitúan dentro de estas grandes tendencias internacionales, y conviene tener en cuenta la experiencia al respecto.

Pros y contras de los sistemas de rendición de cuentas basados sólo en pruebas

En el prefacio de una importante obra sobre el tema, Hamilton, Stecher y Klein (2002), plantean con claridad la idea básica que sustenta el establecimiento de sistemas de rendición de cuentas basados únicamente en pruebas:

> *Los sistemas de rendición de cuentas basados en pruebas se basan en la creencia de que la educación pública puede mejorar gracias a una estrategia sencilla: haga que todos los alumnos presenten pruebas estandarizadas, y asocie consecuencias fuertes a las pruebas, en la forma de premios cuando los resultados suben y sanciones cuando no ocurra así*. (2002, p. iii)

El conjunto de la obra se dedica a analizar la cuestión, mostrando que es mucho más compleja. Pueden destacarse las siguientes observaciones:

- No es fácil hacer pruebas estandarizadas de alta calidad, válidas, confiables y justas. Dada la amplitud de los contenidos curriculares y la dificultad de evaluar en forma estandarizada el rendimiento en los aspectos más complejos, las pruebas pueden limitarse a algunos tópicos, sin duda importantes, dejando fuera otros, incluyendo varios que deberían atenderse con prioridad, como los niveles cognitivos más altos y también los aspectos de actitudes y habilidades.
- Si los resultados tienen consecuencias fuertes para los alumnos y también para las escuelas, los maestros y las autoridades escolares, tenderán a modificar el tiempo dedicado a la enseñanza, priorizando los temas cubiertos por las pruebas; si éstas dejan fuera aspectos importantes del currículo, el resultado será negativo, aun si los puntajes en las pruebas suben.
- Para que lo anterior no ocurra, es necesario que haya una cuidadosa alineación de pruebas y currículo, con base en estándares de contenido y desempeño bien definidos, lo que no es sencillo.
- Los resultados bajos de ciertos alumnos en las pruebas, pueden deberse, en cierta medida, a deficiencias derivadas de sus circunstancias familiares y sociales, y no a falta de capacidad o esfuerzo; se deberá tener mucho cuidado para evitar que se tomen medidas correctivas contrarias a la equidad.
- Los bajos resultados se pueden deber también, en parte, a deficiencias de la propia escuela; algunas de ellas podrán corregirse con medidas administrativas que es relativamente sencillo definir, pero otras requerirán esfuerzos importantes y recursos consi-

derables; es importante que las autoridades tengan conciencia de que el mejoramiento educativo no resultará automáticamente de la evaluación.

- Aun si las pruebas están bien hechas y se aplican correctamente, la decisión de hacerlo en forma censal o empleando muestras de alumnos, áreas curriculares y grados, es inevitable en la mayoría de los casos. Las consecuencias de ello deben tenerse en cuenta, en lo que se refiere a la interpretación de los resultados y el alcance que se les podrá dar.
- Inclusive en el caso de pruebas aplicadas censalmente, hay efectos no muy bien conocidos pero innegables de inflación o de volatilidad de resultados, que hacen aconsejable la máxima prudencia en la interpretación de los resultados de una sola aplicación. Es clara la conveniencia de considerar los promedios de varias aplicaciones para detectar con mayor confiabilidad las tendencias reales.
- La forma de analizar los resultados de las pruebas, en particular la interpretación de los puntajes con referencia a norma o a criterio, tiene pros y contras, Por ello convendrá emplear varios tipos de análisis, además de buscar que los usuarios tengan suficientes elementos para interpretar correctamente los resultados.
- Particular atención debe prestarse al establecimiento de puntos de corte, que inevitablemente implica juicios subjetivos de expertos, así como a la eventual integración de diversos indicadores en un índice global de calidad.
- En general, deberá buscarse la mayor calidad técnica posible, incorporando los avances metodológicos y desarrollando nuevos modelos de evaluación, acordes con dichos avances; se consideran en particular, la Teoría de Respuesta al Reactivo y la psicología cognitiva.
- Por otra parte, se recomienda buscar que se atiendan equilibradamente los propósitos de rendición de cuentas y de mejoramiento de la enseñanza, integrando las perspectivas profesional y política.

En relación con esta última idea, la obra habla de *reconciliar los imperativos de la toma de decisiones de política con buenos estándares técnicos de los sistemas de pruebas*, y hace las siguientes recomendaciones a los tomadores de decisiones:

Los tomadores de decisiones deben considerar todos los costos de los sistemas de pruebas que pretenden implementar. El más importante de dichos costos es el que resulta de la necesidad de ofrecer, a cada alumno que será sometido a pruebas de consecuencias fuertes, las oportunidades adecuadas y apropiadas de aprender el contenido de las pruebas... Los tomadores de decisiones necesitan también persuadir a sus representados de que sean más pacientes en sus juicios sobre la educación pública... Persuadir al público de ser paciente exigirá evidencias de que las escuelas, sin duda, están respondiendo a las expectativas públicas.

Pero otro aspecto de esta promoción de la paciencia por parte del público consiste en convencer a los ciudadanos de que la rendición de cuentas es un asunto de ida y vuelta: las

escuelas no podrán cumplir con los estándares que la comunidad les fije, si la comunidad misma no cumple con su obligación de apoyar adecuadamente a las escuelas. Siempre será difícil convencer de que no hay soluciones fáciles a los problemas, pero el éxito que han tenido algunos líderes políticos sugiere que es posible lograrlo. Cerrar la brecha que existe entre los imperativos políticos y los estándares psicométricos implica, sobre todo, que los tomadores de decisiones acepten las limitaciones de las pruebas y de sus usos potenciales... (Hamilton, Stetcher y Klein, 2002, pp. 117-118)

A las consideraciones anteriores deben añadirse otras, derivadas de la naturaleza multidimensional del concepto de calidad educativa. Como se indicó en el primer apartado de este documento, la calidad no debe identificarse sin más con el nivel promedio de aprendizaje alcanzado por los alumnos de una escuela o sistema educativo. Hay que tener en cuenta otras dimensiones de la calidad, en particular la cobertura y la eficiencia terminal, sin olvidar la eficiencia (costo) y la equidad.

En efecto: si los alumnos de una escuela obtienen resultados elevados en ciertas pruebas de rendimiento, ello puede no ser el resultado del trabajo de la escuela, sino de otros factores, unos que no dependen de la escuela, como es el nivel socioeconómico de los alumnos, pero también otros que sí dependen de decisiones escolares.

Si una escuela, por ejemplo, selecciona de manera estricta a su alumnado, quedándose con los mejores aspirantes y dejando fuera a los más débiles, los resultados que los aceptados obtendrán en pruebas posteriores serán, probablemente, mejores que los de otro plantel que deba o decida admitir, tanto a los aspirantes de buen nivel como a los menos buenos. De manera similar, una escuela que permita que los alumnos más débiles deserten, o incluso los orille a hacerlo, conservando sólo a los mejores, obtendrá mejores resultados, que otra que conserve hasta el final del trayecto escolar a todos los alumnos que hayan ingresado a ella en un momento dado.

No es sencillo decidir cuál escuela es mejor: una selectiva al ingreso y con alta deserción, cuyos egresados tengan altos resultados, u otra que admita a todo aspirante y no tenga deserción, pero cuyos resultados promedio sean inferiores a la anterior.

Los resultados de las pruebas PISA difundidos el 1 de julio de 2003, nos ofrecen un caso adecuado para ejemplificar este tipo de situaciones.

Una lectura superficial de los resultados obtenidos por los alumnos mexicanos en comparación con los alcanzados por los de Argentina o Chile, puede dar lugar a la afirmación simple de que el nivel educativo de México es mejor, así sea ligeramente, al de los otros dos países. Pero si se observa que en México, la proporción de jóvenes de 15 años que están inscritos en secundaria o bachillerato (lo que constituye la población objetivo de las pruebas de PISA), representa poco más del 51 por ciento, en tanto que en Argentina es de 76 y en Chile llega a 87 por ciento. Resulta claro que una afirmación simplista como la mencionada es inadecuada: si en México también siguiera en la escuela el 76 o el 87 por ciento de

los jóvenes de 15 años, muy probablemente el promedio de los resultados sería inferior al obtenido, ya que los desertores son, en su mayoría, alumnos de bajo rendimiento.

Para tener un buen sistema de rendición de cuentas educativas no basta, por lo tanto, tener información sobre los resultados obtenidos por los alumnos en pruebas de rendimiento: hacen falta datos de cobertura, deserción, etcétera.

Diferencias de los sistemas de evaluación orientados únicamente a la rendición de cuentas y los orientados al mejoramiento

El sistema educativo de Texas es citado frecuentemente en forma elogiosa por sus recientes avances, que se atribuyen a la introducción de sistemas de rendición de cuentas basados en pruebas, y a las medidas derivadas del análisis de sus resultados. La adopción de políticas federales en este sentido, mediante la ley *No Child Left Behind*, es muestra de las expectativas que se depositan en este tipo de sistemas, impulsados por el actual presidente Bush desde su gestión como gobernador de Texas, y ahora al frente de la administración federal americana.

Sin embargo, un análisis más cuidadoso, muestra otro ángulo del problema: la extensión del uso de pruebas de rendimiento en el sistema escolar texano, se remonta a la década de los ochenta; pero en la siguiente se dio un desarrollo muy interesante, en la forma de un sistema que complementa los resultados obtenidos por los alumnos en las pruebas de aprendizaje, con informaciones sobre las características de las escuelas y su entorno y, sobre todo, al utilizar los resultados de las evaluaciones de manera sistemática para el diseño de acciones de mejoramiento.

Esto incluye un sistema de difusión de resultados, que contextualiza los puntajes obtenidos en las pruebas por los alumnos, añadiendo información sobre variables clave del alumno y su entorno. Con esto es posible formar grupos de escuelas comparables según las variables de los alumnos y su entorno. De esa forma, pueden detectarse no sólo aquellas escuelas que obtienen los resultados más altos en términos absolutos, sino también las que destacan por sus resultados *en relación con escuelas comparables*. Esta forma de difundir y utilizar los resultados resulta más equitativa y aceptable a los maestros, pero además permite impulsar acciones de mejora, a partir de las prácticas de las mejores escuelas de los diversos contextos socioeconómicos y culturales.

La organización no gubernamental *Just for the Kids,* ha jugado un importante papel en el desarrollo de un sistema de difusión de resultados de este tipo, que ha sido adoptado ya por más de veinte estados de la Unión Americana, y está extendiéndose a los restantes, gracias a un apoyo especial del gobierno federal.

A continuación se presenta una comparación de un sistema de evaluación enfocado exclusivamente a la rendición de cuentas, en contraste con otro que, sin descuidar lo an-

terior, ponga el énfasis en la contribución a la mejora de las escuelas. Tanto esta comparación como la información sobre las actividades de la organización *Just for the Kids,* fueron proporcionadas por un investigador que ha participado en forma importante en ellas, (Cfr. López, 2003).

Modelos de evaluación centrados en rendición de cuentas	**Modelos de evaluación enfocados a difusión de mejores prácticas**
Se proponen hacer a las escuelas responsables del uso de recursos públicos.	Se proponen mejorar las escuelas a partir de la adopción de las mejores prácticas de las escuelas de altos resultados en grupos de escuelas comparables.
Usan un solo punto de referencia para evaluar: la proporción de alumnos de una escuela, que obtiene resultados por encima de cierto punto de corte.	Usan puntos de referencia múltiples, que tienen en cuenta diversos grupos de alumnos, según sus resultados escolares previos.
Si un alto porcentaje de alumnos obtiene resultados por encima del punto de corte, aunque sea ligeramente, no hay elementos para promover mejora adicional	Al tener en cuenta las diferencias de las escuelas, es más factible establecer procesos de mejora contínua.
No toma en cuenta las condiciones de la escuela ni los factores de los alumnos; los resultados tienden a generalizarse como si todo el alumnado fuera igual. Fomenta sentimiento de inequidad de la evaluación y dificulta la difusión de prácticas interesantes de escuelas de contextos desfavorables.	La consideración de las condiciones de la escuela y las variables de los alumnos, facilita identificar subgrupos de alumnos con necesidades específicas. La evaluación se ve equitativa y se facilita el interés por conocer las prácticas de escuelas comparables de altos resultados.
Mira hacia el pasado, preguntándose cuál ha sido el desempeño de alumnos de una escuela hasta ahora.	Toma en cuenta el desempeño previo, pero además se interesa por el posible desempeño futuro.
Perspectiva del vaso medio vacío: si el promedio es bajo la escuela es mala, aun con buenos alumnos.	Perspectiva del vaso medio lleno: aun en escuelas de resultados promedio bajos hay buenos alumnos.
Evaluación sumativa, fin de ciclo; no retroalimenta al maestro a tiempo para que trate de mejorar.	Puede ser sumativa, pero es mejor si es formativa, a lo largo del curso, con retroalimentación oportuna.
Termina en juicio sobre la efectividad de la escuela para alcanzar el nivel mínimo aceptable.	Permite enfoque de mejora continua hacia niveles más exigentes, a partir de mejores prácticas.
No ofrece apoyo al maestro para decidir qué hacer para mejorar; puede llevar a usar sanciones como medio principal de mejora.	Por el intercambio de información sobre mejores prácticas de escuelas comparables, apoya a maestros para adoptar estrategias prometedoras de mejora.
Da por supuesto que los bajos resultados de una escuela, son un problema temporal que será fácil resolver con la simple amenaza de sanciones.	Supone que los bajos resultados son un problema duradero, porque tiene causas complejas, cuya solución requiere plazos largos y estrategias complejas

Si bien la caracterización del cuadro anterior puede ser algo simplista, que caricaturiza los dos modelos de evaluación cargando las tintas oscuras en un lado y las claras en el otro,

las dos casillas del último renglón contienen una observación cuya importancia conviene subrayar: es frecuente que las personas ajenas al medio educativo desconozcan la complejidad de los procesos de enseñanza y aprendizaje. Suele ocurrir, en consecuencia, que se subestime la dificultad de los cambios y las mejoras.

Si hacer una buena enseñanza, con cualquier grupo de alumnos y en cualquier contexto fuera algo sencillo, el que en algunos casos no se haga y los resultados sean bajos, sólo puede ser el resultado de una gran negligencia. Mejorar, por lo tanto, debe ser fácil, con tal que se introduzcan mayores niveles de exigencia. Por el contrario, si se entiende que, además de indudables negligencias, los bajos resultados educativos son el resultado de un complejísimo entramado de factores causales, se entenderá también que la mejora requiere, como se ha apuntado, de tiempos largos y estrategias complejas.

LA TENSIÓN ENTRE LA NECESIDAD DE RESULTADOS POR ESCUELA Y LAS EXIGENCIAS TÉCNICAS

Hay una tensión entre dos elementos importantes: de un lado la legítima demanda de tener resultados por escuela –y aun por alumno– para que maestros, alumnos y padres de familia tengan la retroalimentación necesaria para ajustar sus respectivas acciones; del otro, la imposibilidad técnica, al menos en sistemas educativos grandes, de hacer evaluaciones de calidad controlada, para asegurar su confiabilidad y comparabilidad, de todos los alumnos y escuelas, en todos los grados y cubriendo adecuadamente todos los aspectos del currículo.

La manera en que tratan de atender las dos partes de este dilema los sistemas de evaluación de algunos países, que se consideran de los mejores en su campo, muestra rasgos interesantes, que apuntan en una misma dirección.

- Las evaluaciones censales, se presentan más fácilmente en sistemas educativos de dimensiones reducidas, como los de Singapur o Uruguay. Conviene recordar que el sistema educativo mexicano tiene cerca de 15 millones de alumnos en casi 100 mil escuelas primarias; Holanda, con una población de 15.9 millones, no mucho mayor que la del estado de México, y una demografía más madura, tiene sólo 1.6 millones de alumnos en más de 7 mil escuelas del nivel equivalente. Chile tiene una población y un sistema educativo de dimensiones comparables. Uruguay y Singapur, con 3.5 y 4.2 millones de habitantes, tienen menos alumnos en primaria que la capital de Jalisco.
- En sistemas grandes, como el de Estados Unidos –con unos 24 millones de alumnos en el nivel equivalente a la primaria– la evaluación censal corre a cargo de las entidades federativas, mientras que a nivel nacional se maneja un sistema muestral. Los sistemas estatales ofrecen resultados por escuela, lo que es posible por su carácter censal; el NAEP, en cambio, no lo hace. Según la legislación estadounidense, publicar los resultados del

NAEP por escuela constituiría un delito. De hecho tal cosa nunca ha ocurrido en los más de treinta años de existencia del sistema, (Cfr. NAEP, 2002).

- La institución holandesa responsable de la evaluación educativa (conocida por las siglas CITO), probablemente el organismo de evaluación más reconocido de Europa por su calidad técnica, combina también dos acercamientos: por una parte, pruebas masivas, que se aplican a todos los alumnos en cada escuela, sin control externo, y retroalimentan a alumnos, padres y maestros de inmediato; por otra, evaluaciones muestrales controladas, cuyos resultados no se difunden por escuela ni sirven para decisiones sobre personas o planteles singulares, (Cfr. Forster y Valverde, 2003).
- Otro país líder en este campo es Australia, que cuenta también con sistemas dobles: en la provincia de Australia Occidental, por ejemplo, se manejan dos sistemas paralelos: uno para monitorear las tendencias a nivel nacional, basado en muestras, y otro de evaluación de escuelas individuales: MSE y WALNA. El sistema nacional por muestreo no ofrece resultados por escuela, (Cfr. Forster y Valverde, 2003).
- Las conclusiones de la revisión del sistema chileno de evaluación (Sistema de Medición de la Calidad Educativa, Simce), recientemente difundidas incluyen la recomendación de mantener la difusión de resultados por escuela, como permiten las evaluaciones censales en grados clave llevadas a cabo desde 1988, pero recomiendan también que se complementen con evaluaciones muestrales, las cuales permitan valorar con mayor precisión las tendencias en el tiempo, además de extenderse a nuevas áreas curriculares y grados, (Cfr. Comisión para el Desarrollo y Uso del Simce, 2003).
- También es interesante la experiencia del sistema inglés, el más conocido por difundir resultados por escuela con consecuencias potenciales fuertes para cada establecimiento. Como se ha apuntado, el sistema del Reino Unido comenzó, a partir de 2003, a difundir los resultados con base ya no en puntajes crudos, sino en mediciones del valor agregado por cada plantel. Esta medida es el resultado de la creciente conciencia de que las evaluaciones de escuelas basadas en resultados simples pueden llevar a conclusiones injustas. Aun en un sistema tan fuerte, que tiene más de medio siglo de hacer evaluaciones educativas en gran escala, y más de una década de difundir resultados por escuela, se reconoce que la metodología actualmente utilizada para estimar el valor agregado es pobre.

De nuevo sobre las diferencias entre evaluaciones, según sus consecuencias

Tras el repaso de experiencias de diversos sistemas educativos, en lo relativo a difusión de resultados de evaluación por escuela, conviene regresar al punto de partida: el de las diferencias entre las evaluaciones que se proponen monitorear tendencias en gran escala,

y las que buscan retroalimentar directamente la práctica docente. Las consecuencias que se derivarán de una evaluación determinan, en gran medida, las metodologías apropiadas para hacerla, y la forma de difundir y usar sus resultados.

Son consideradas de *alto impacto* las evaluaciones que llevan a decisiones importantes para el individuo, o la institución que se somete a la evaluación, como aprobar o reprobar; ser admitido en un programa o graduarse de él; recibir o no la acreditación para ofrecer determinados estudios. De *bajo impacto* son las evaluaciones que no llevan a decisiones de tales consecuencias para las personas o instituciones en lo individual. En este rubro cabe la evaluación llamada formativa, la cual retroalimenta al evaluado para que pueda mejorar su desempeño, sin llevar de inmediato a decisiones. Pueden incluirse también evaluaciones que llevan a la formulación de diagnósticos para sustentar decisiones de políticas en el nivel macro, pero no sobre personas o instituciones en lo individual.

Por la gravedad potencial de sus consecuencias, las evaluaciones individuales de alto impacto deben basarse en la evidencia más amplia de que pueda disponerse. Por ello, las evaluaciones de aprendizaje en gran escala con instrumentos estandarizados (con preguntas que constituyen muestras relativamente pequeñas de los dominios a evaluar), no pueden ofrecer un sustento suficiente para decisiones tan delicadas, tampoco es esperable que lo puedan hacer en un futuro previsible, aunque en teoría sea posible.

Para que decisiones individuales importantes pudieran basarse sólo en los resultados de pruebas aplicadas en gran escala, éstas deberían aplicarse a todos los alumnos; ser de una extensión mucho mayor a la habitual, para cubrir todas las áreas del currículo y todos los temas de cada área; aplicarse no sólo al fin del ciclo escolar sino también al inicio del mismo, e incluso en momentos intermedios; obtener resultados en plazos muy cortos, para que el efecto positivo de retroalimentación ocurra.

Esto es posible para otorgar licencias para el ejercicio de profesiones en que los evaluandos no son numerosos y las competencias mínimas a dominar pueden precisarse con toda claridad. En estos casos, además, las consecuencias para la sociedad de decisiones erróneas derivadas de una evaluación deficiente, pueden ser muy serias, por lo que se justifican aplicaciones de larga duración, adecuadas también a la edad de los sustentantes.

Características opuestas hacen impensable algo similar en educación básica. Si bien teóricamente podría pensarse en un sistema de evaluación que tuviera cobertura completa, tanto del universo de alumnos como de los contenidos, eso es imposible cuando se trata de millones de alumnos. Las pruebas en gran escala no pueden ser la única base de decisiones sobre cada alumno. Es indispensable tomar en cuenta información más completa sobre cada uno, que sólo puede ofrecer el juicio calificado del maestro. Algo similar puede decirse en cuanto a la evaluación de las escuelas como tales, o de los maestros en lo individual. El nivel de aprendizaje alcanzado por los alumnos, es sólo una de las dimensiones a tener en cuenta en un juicio sobre la calidad de unas y otros,

además de que depende no sólo de su acción, sino de múltiples factores del entorno social y familiar de los alumnos.

Por las características de las pruebas en gran escala, un juicio sólido sobre la calidad de una escuela o un maestro, no deberá basarse en los resultados de una aplicación. Serían necesarias varias aplicaciones para evitar la volatilidad de los resultados y poder valorar las tendencias. Para contar con medidas del valor agregado por una escuela o docente, se necesitan mediciones de entrada y salida o de inicio y fin de ciclo. Se necesita además controlar los factores de contexto y valorar los procesos del interior de la escuela o el aula.

Un juicio sobre calidad de escuelas y maestros en lo individual, requiere información que no puede derivarse sólo de pruebas de aprendizaje en gran escala; implica procesos prolongados de observación directa, normalmente a cargo de directores y supervisores. No debe olvidarse que, en este terreno, hay también un trabajo importante de capacitación por realizar, ya que la forma en que se realiza la función de dirección y la de supervisión, dista muchas veces de ser satisfactoria. El papel de las evaluaciones en gran escala, en cambio, es fundamental para llegar a juicios sobre la situación de un sistema educativo como tal, diagnósticos que sirvan para sustentar decisiones de política en el nivel macrosocial.

Las evaluaciones con propósitos de mejora pedagógica en escala micro, por la exigencia de retroalimentación inmediata a maestros, alumnos y padres de familia, estarán preferentemente a cargo de instancias cercanas al aula. En sistemas grandes, y como estas evaluaciones no deben usarse para comparaciones en gran escala, no es necesario que su aplicación sea controlada externamente, éstas pueden estar a cargo de cada escuela, al cuidado de directores y supervisores, con participación de los padres de familia. Para que docentes y escuelas puedan realizar bien este tipo de evaluaciones, es necesario dotarlos de orientaciones e instrumentos adecuados. Esas orientaciones no deberán limitarse a informar sobre la escala de calificaciones a utilizar.

Establecer si un alumno tiene o no competencia en ciertas partes del currículo, es tarea delicada y técnicamente compleja, tanto en lo relativo a la manera de evaluar, como en cuanto a la manera de usar la evaluación para apoyar el aprendizaje de los alumnos y en cuanto a la forma de informar a las familias sobre el avance de cada uno, las dificultades que enfrenta y la forma de apoyarlo.

Las evaluaciones con propósitos de diagnóstico para decisiones de mejora educativa en gran escala, en cambio, deberán llevarse a cabo mediante procesos de aplicación cuidadosamente controlados, preferentemente por una instancia externa, para asegurar la ausencia de sesgos y la confiabilidad de los resultados.

Las evaluaciones pueden complementarse, si se articulan adecuadamente: las evaluaciones muestrales y controladas que se aplican en gran escala, permiten hacer juicios sólidos sobre el conjunto del sistema educativo y sus grandes subsistemas, que se pueden comparar entre sí y, en su caso, con los resultados de pruebas internacionales, lo que eva-

luaciones no controladas no pueden ofrecer. Esto es importante para la toma de decisiones de política educativa. La evaluación desarrollada a nivel local, además de ofrecer elementos clave para la mejora pedagógica, permitirá a escuelas y maestros compararse con los estándares nacionales, si emplea instrumentos de calidad similares a los que se utilizan en las aplicaciones controladas.

Es importante, desde luego, que todas las evaluaciones sean de buena calidad, que tengan los mismos referentes curriculares, utilicen metodologías consistentes y apliquen estándares de calidad semejantes, para que los resultados de las aplicaciones masivas puedan entenderse a la luz de los que arrojen las muestrales controladas.

No es sencillo lograr una buena conexión entre evaluaciones muestrales controladas, evaluaciones en gran escala aplicadas por las escuelas y las que hacen los maestros en el aula. La novedad de los dos primeros enfoques, hace que no haya experiencia sobre el particular. Por ello es importante emprender de inmediato el desarrollo de modelos para asegurar esa conexión. Se debe añadir que la participación de los maestros en las evaluaciones en gran escala, puede contribuir considerablemente a su desarrollo profesional, mejorando su capacidad para hacer buenas evaluaciones en la escala del aula.

CONCLUSIONES

Las consideraciones presentadas en las páginas anteriores llevan a la conclusión de que es fundamental tener en cuenta las características técnicas de una evaluación, para decidir cuáles usos de la misma son legítimos y cuáles no lo son. Para que los resultados de los puntajes de los alumnos de una escuela, puedan ser utilizados para valorar la calidad del plantel en donde estudian, deberían reunirse una serie de condiciones, que se resumen como sigue:

- Las pruebas deberían haberse aplicado en todas las escuelas del sistema de que se trate y no sólo en una muestra de ellas. No sólo se necesita censar escuelas, sino también grupos y alumnos, ya que puede haber diferencias importantes en el nivel de aprendizaje de un grupo a otro.
- Dada la importancia que tiene el contexto socioeconómico de los alumnos en sus resultados en pruebas, éstos deberán ir acompañados por información sobre su entorno para que las comparaciones lo tengan en cuenta, cuidando la equidad. Por lo mismo, no deberían utilizarse puntajes crudos, sino las diferencias entre los puntajes obtenidos y los deseables en función del entorno.
- Como la diferencia de los puntajes obtenidos por dos escuelas en una prueba puede ser o no significativa, lo cual depende del tamaño del error estándar, siempre deberá darse la información suficiente para que se pueda apreciar si cierta diferencia es o no significativa.

- Como puede haber cambios significativos en el nivel de aprendizaje de los alumnos de una misma escuela, de un año a otro, tanto por razones de la propia escuela como del entorno, los juicios sobre la calidad de una escuela no deberán sustentarse en los resultados de una sola aplicación, sino en la tendencia a lo largo de varios años. Lo deseable sería tener medidas de la ganancia alcanzada por los alumnos a lo largo del tiempo, lo que implicaría estudios longitudinales, con mediciones en varios puntos del tiempo y seguimiento individual de estudiantes. Deberá valorarse también la permanencia de los alumnos en la escuela de que se trate.
- Además de los resultados del aprendizaje, hay otras cualidades importantes de una buena escuela, como el estado físico de las instalaciones; la existencia de materiales de apoyo; el ambiente de orden y trabajo; la dedicación de los maestros; el trato que den a los alumnos; el liderazgo del director; la participación de los padres de familia; etcétera. Por ello, los datos sobre los puntajes obtenidos por los alumnos en pruebas, deberían ir acompañados también por información de esos aspectos, tratando de distinguir los efectos escolares de los del contexto.

En el marco de una política de transparencia y de búsqueda de una participación más importante y comprometida, de los padres de familia y la sociedad entera, en los esfuerzos de mejora, no puede minimizarse la importancia de las razones a favor de la difusión de resultados de las evaluaciones por escuela.

Tampoco se puede ignorar el peso de las experiencias que muestran los efectos negativos de una difusión descontextualizada de los resultados de las pruebas, ni la fuerza de los argumentos, que llaman la atención sobre el cuidado que se debe tener para no usar las evaluaciones en una forma que no corresponda a sus características técnicas.

Los organismos responsables de hacer evaluaciones en gran escala, además de asegurar la calidad técnica de su trabajo, deberán cuidar la forma de difundir y usar los resultados, evitando atentar contra la equidad y buscando maximizar el efecto positivo de las evaluaciones para la mejora de la calidad educativa. Al difundir resultados por escuela, siempre deberá darse información que permita a los usuarios interpretarlos correctamente, teniendo conciencia de sus alcances y límites.

Los organismos responsables de evaluaciones en gran escala, no deberán dar resultados por escuela, de manera congruente con las prácticas de sistemas de evaluación en gran escala, tan importantes como el NAEP de los Estados Unidos.

Los organismos mencionados, podrán apoyar a las autoridades educativas para que lo hagan, siempre y cuando las evaluaciones a difundir reúnan las condiciones indispensables para que se eviten las distorsiones más riesgosas, como el que se trate de evaluaciones censales, que incluyan información del contexto socioeconómico y que se basen en resultados de al menos tres ciclos escolares. Los sistemas de difusión de resultados por escuela deberían tratar, además, de incorporar progresivamente otros elementos, como medidas

del valor agregado por la escuela y de la ganancia que muestre el aprendizaje de los alumnos en el tiempo.

REFERENCIAS BIBLIOGRÁFICAS

COMISIÓN PARA EL DESARROLLO Y USO DEL SIMCE (2003). *Evaluación de Aprendizajes para una Educación de Calidad*. Santiago de Chile. Ministerio de Educación.

FORSTER, MARGARET y GILBERT A. VALVERDE (2003). *La Experiencia Internacional en Sistemas de Medición: Estudios de Casos*. Santiago de Chile. Ministerio de Educación.

GONZÁLEZ, PABLO (2003). *Evaluar para aprender o para deprimirse*. Citado por Elacqua, Gregory. Universidad Adolfo Ibañez. Escritorio de investigadores. http://www.educarchile.cl/modulos/noticias/constructor/investigador.asp

HAMILTON, LAURA S., BRIAN M. STECHER y STEPHEN P. KLEIN Eds. (2002). *Making Sense of Test-Based Accountability in Education*. Santa Monica CA, Rand Education.

HAMILTON, LAURA S. Y DANIEL M. KORETZ (2002). Tests and Their Use in Test-Based Accountability Systems. En HAMILTON, STECHER y KLEIN, 2002, PP. 13-49.

LÓPEZ, OMAR (2003). *Comunicación personal sobre Just for the Kids*.

SCHIEFELBEIN, ERNESTO (2003). "SIMCE, ¿castigar o ayudar?". *La Tercera*. Santiago de Chile. Junio.