Evaluación e investigación educativa

Felipe Martínez Rizo y Annette Santos del Real

Introducción

Un rasgo que destaca en el panorama educativo mexicano reciente, es la creciente visibilidad e importancia de la evaluación educativa y, en particular, de la modalidad que se basa en la aplicación en gran escala de instrumentos estandarizados, que buscan medir el nivel de aprendizaje que alcanzan los alumnos en ciertas áreas.

Además de la influencia de tendencias similares en el plano internacional, la creciente importancia que se atribuye a la evaluación se relaciona con una preocupación, también creciente, por la calidad de la educación. En un gran número de países, con niveles de ingresos altos y bajos, la idea de que una educación de buena calidad es un componente fundamental de toda estrategia de desarrollo es ampliamente compartida por líderes políticos, dirigentes empresariales, educadores y el público en general. Sin creer que la educación tiene las virtudes casi mágicas que a veces se le atribuyen retóricamente, parece sólidamente sustentada la opinión de que tanto una economía competitiva, como una democracia moderna y, en general, una sociedad avanzada, suponen una ciudadanía con niveles considerables de escolaridad, de buena calidad.

En países de desarrollo intermedio como México, la participación en evaluaciones internacionales ha dado evidencias duras de que, además de menores cifras de escolaridad y cobertura, los niveles de aprendizaje que alcanzan los alumnos son también inferiores a los que consiguen los estudiantes de países más avanzados. Estos datos, que no deberían sorprender a nadie, han hecho que el tema de la calidad educativa, así como el de su evaluación, se pongan de moda; es frecuente, sin embargo, que estas ideas se manejen de manera simplista, tanto en lo que se refiere a la descripción de la situación, como en lo que toca a sus causas, y también a la forma de valorarla y corregirla.

Las visiones simplistas de la situación llevan fácilmente a afirmaciones excesivamente severas, sin referentes que permitan formular juicios equilibrados. Esos diagnósticos, además, no toman en cuenta la complejidad de los factores de las escuelas y del entorno familiar y social que influyen en los resultados. La falta de contextualización de los diagnósticos, a su vez, lleva a proponer estrategias simples para remediar la situación que, según la inclinación ideológica de quienes las formulen, se reducen a proponer que se privaticen las escuelas públicas, o bien a que se incrementen los recursos públicos que se les asignan. En la misma tónica, las posturas simples tienden a pensar que los medios para evaluar la calidad se reducen a la aplicación de pruebas de rendimiento, sin distinguir su enfoque y nivel técnico.

La consecuencia es una proliferación de pruebas para valorar la calidad de las escuelas, aunque no hayan sido diseñadas para tal propósito y aunque adolezcan de serias deficiencias técnicas. Aparecen entonces dos tipos de reacciones: los maestros se ven impulsados a orientar su trabajo pedagógico en función de los contenidos de las pruebas, ya que serán el medio privilegiado para valorar su trabajo; por otra parte, entre los mismos maestros y entre los estudiosos de la educación, es frecuente un rechazo absoluto a las pruebas, por considerarlas intrínsecamente inadecuadas para valorar correctamente la calidad.

Los autores de este trabajo pensamos que el amplio acuerdo que parece haber en cuanto a la idea de que la evaluación es importante, no debe hacer perder de vista que no cualquiera es positiva; que mal concebida o realizada, la evaluación puede ser irrelevante, en el mejor de los casos, o incluso destructiva. Los sistemas educativos no necesitan evaluaciones de cualquier tipo, aunque sean destructivas, sino unas adecuadas, que contribuyan efectivamente a que la calidad educativa mejore.

de Alba Alicia (Ed.). ¿Qué dice la investigación educativa? México. COMIE, pp. 265-304.

Con base en lo que hemos aprendido a lo largo de los años en que nos hemos dedicado a la evaluación educativa, y aprovechando varios textos previamente publicados por el Instituto Nacional para la Evaluación de la Educación (INEE), en este artículo pretendemos ofrecer una visión más completa sobre dicha actividad, de cuya importancia potencial para el sistema educativo nacional estamos convencidos, pero de cuyos posibles repercusiones negativas somos también conscientes.

Antes de entrar en materia es importante hacer una distinción entre la evaluación misma, que consideramos como un tipo de investigación que puede contribuir a enriquecer el conocimiento que se tiene sobre la realidad educativa, y la investigación de segundo nivel que se puede hacer sobre la evaluación misma.

Desde principios de la década de 1990, y con mayor fuerza en la primera del siglo XXI, ha habido en México avances considerables de la investigación evaluativa, pero la investigación sobre la evaluación no ha avanzado al mismo ritmo. Por ello el contenido de este capítulo se organiza de manera que la mayor parte del texto se dedica a revisar qué nos dice la investigación educativa de tipo evaluativo sobre la educación mexicana; un segundo apartado, de menor extensión, se refiere a la investigación sobre la evaluación misma, apuntando principalmente lo que sería deseable tener y que aún no se ha hecho.

Por lo que se refiere al primer punto, después de un breve repaso de los orígenes de la evaluación educativa, y en especial en su modalidad de pruebas estandarizadas de aprendizaje, se presentan algunas ideas básicas en relación con el diseño de tales instrumentos; se describen luego las principales evaluaciones en gran escala que actualmente se aplican en México, señalando sus propósitos, alcances y límites.

La conclusión del capítulo se dedica a una reflexión sobre el papel que, a nuestro juicio, debería tener la evaluación en gran escala en el sistema educativo mexicano, para que sea realmente un factor que contribuya al mejoramiento de la calidad; en este punto se destaca la relación que debería haber entre la evaluación en gran escala y la evaluación que cada maestro lleva a cabo en el aula.

Conviene subrayar que la evaluación educativa no debe reducirse a la aplicación de pruebas de aprendizaje, sino que debe incluir otros acercamientos a otras dimensiones de la calidad educativa, como la cobertura del sistema, su eficiencia y equidad, su impacto en la vida adulta, etc. No sobra añadir que este texto refiere fundamentalmente a evaluaciones del aprendizaje en educación básica.

1. ¿Qué nos dice la investigación evaluativa sobre la educación mexicana?

1.1. Antecedentes de la evaluación de aprendizajes en gran escala

Tradicionalmente la evaluación ha formado parte de las actividades que se llevan a cabo en las escuelas. La manera en que se evaluaba, desde luego, era muy diferente de la que se utiliza hoy en las evaluaciones en gran escala. Los maestros estimaban el grado en que cada alumno conseguía adquirir los conocimientos y habilidades que se pretendía desarrollara, y lo hacían mediante estrategias estrechamente relacionadas con la actividad docente, de la que evaluar era una parte.

Esos procedimientos, hoy todavía vigentes, incluían las preguntas que los alumnos debían responder de viva voz; la valoración de su capacidad lectora escuchándolos leer en voz alta pasajes de los libros; la observación de su desempeño cotidiano, individual o en grupo, por ejemplo pasando al pizarrón, o revisando las tareas que debían hacerse en casa.

Cuando las escuelas atendían sólo a una minoría de los niños de cada generación, muchos de hogares privilegiados, como ocurría en México a principios del siglo XX, esos procedimientos tradicionales bastaban para asegurar que todo alumno alcanzara un mínimo de conocimientos y habilidades. En cambio, en la medida en que llegaron a

las escuelas niños de sectores menos favorecidos, con la expansión del sistema educativo a zonas cada vez más marginadas, aumentó también la heterogeneidad de su nivel de rendimiento, y los estándares de evaluación implícitos manejados por los docentes se diversificaron en función del contexto.

Las metodologías de enseñanza, sin embargo, no se diversificaron en la misma medida para atender las necesidades de cada grupo de alumnos, y los recursos que se dieron a las escuelas, en lugar de jugar un papel compensatorio o de discriminación positiva, contribuyeron a perpetuar las desigualdades, al ser menores para las escuelas que atienden a los alumnos que tienen condiciones más desfavorables para el aprendizaje en el hogar. Por ello ahora no se puede dar por hecho que terminar cierto grado escolar, o incluso la educación básica, asegure los niveles que se suelen considerar como los mínimos aceptables.

Surgió así la necesidad de desarrollar procedimientos que permitieran medir el aprovechamiento escolar en forma comparable en contextos diversos. Y como la idea de un sistema educativo que atendiera masivamente a todos los sectores de la población se desarrolló en los Estados Unidos antes que en la mayor parte del mundo, no es sorprendente que las evaluaciones en gran escala se hayan desarrollado también en ese país antes que en otros, desde mediados del siglo XIX.

Al parecer, la primera ocasión en que se usaron pruebas estandarizadas para dar cuenta del aprendizaje de los estudiantes fue en 1845, en escuelas de Boston, en las que se evaluaron de esa forma los conocimientos de historia de 500 alumnos. A partir de 1895 se registra en Estados Unidos el uso de pruebas de ortografía, lectura y aritmética, a números de alumnos que iban de 8,300 a 16,000, por parte de J. M. Rice, lo que continuó a principios del siglo XX por el *Nacional Council of Education* y la *National Education Association*. (De Landsheere, 1986)

Durante la primera mitad del siglo XX se desarrolló la tecnología para construir instrumentos de medición que permitan obtener resultados válidos y confiables, dando lugar a la Teoría Clásica de las Pruebas (Cfr. Martínez Arias, 1995). La segunda mitad del siglo pasado y lo que va del presente han visto nuevos y variados desarrollos, como los de las pruebas con referencia a un dominio o criterio, los modelos de respuesta al reactivo, la teoría de la generalizabilidad, las técnicas de equiparación, los diseños matriciales, el uso de preguntas de respuesta construida, las pruebas adaptativas por computadora, entre otros avances.

En paralelo, el uso de pruebas estandarizadas para evaluar aprendizajes en gran escala se ha extendido, hasta llegar a la situación actual, en la son escasos los países en los que no se hacen esfuerzos en este sentido. También desde principios de la segunda mitad del siglo XX surgieron las evaluaciones en gran escala en el plano internacional, con el desarrollo de las primeras evaluaciones de matemáticas y luego de ciencias y otras áreas por parte de la *International Association for the Assessment of Educational Achievement* (IEA).

En nuestro país, como en la mayor parte de los menos desarrollados, el desarrollo fue más lento. Con algunos antecedentes en educación superior en décadas anteriores, fue sobre todo desde la de 1990 cuando este tipo de evaluaciones comenzó a cobrar fuerza en el sistema educativo mexicano, con las pruebas del Factor Aprovechamiento Escolar del programa de estímulos de Carrera Magisterial y las evaluaciones de los programas compensatorios apoyados por el Banco Mundial. Por la misma época comenzó también la participación del país en evaluaciones internacionales, con el primer estudio del Laboratorio Latinoamericano de Evaluación de la Calidad Educativa (LLECE, 1996) y, en forma incompleta, en el Tercer Estudio Internacional de Matemáticas y Ciencias de la IEA (TIMSS, 1995), y luego sobre todo con las pruebas PISA (*Programme for International Student Assessment*), que la OCDE, decidió emprender en 1997, y cuya primera aplicación tuvo lugar en 2000.

1.2. Sentido, propósitos y principios de las evaluaciones en gran escala

Para aprovechar el potencial positivo de las evaluaciones estandarizadas y evitar sus posibles efectos negativos, es obligado comprender que se diseñan con propósitos distintos, y que la atención de esos propósitos exige cualidades técnicas particulares. Siguiendo a Ravela (2006) pueden identificarse al menos seis propósitos:

Evaluaciones para acreditación y/o certificación. El propósito es otorgar algún tipo de constancia formal de que un individuo o institución posee ciertas características o cualidades. Ejemplos de evaluaciones con este tipo de propósito son las que buscan:

- Identificar qué estudiantes han logrado los conocimientos y competencias estipulados al finalizar un grado o ciclo, para otorgarles el certificado correspondiente o autorizar su tránsito al siguiente nivel.
- Certificar qué individuos poseen las competencias necesarias para ocupar cargos de dirección, de supervisión o técnicos, en el sector público o privado, como al concursar plazas directivas o del servicio civil de carrera.
- Acreditar instituciones o programas educativos, por ejemplo dando el reconocimiento oficial a universidades o carreras.

Evaluaciones para ordenar individuos o instituciones. En este caso no se busca llegar a juicios absolutos en relación con ciertos parámetros, sino a juicios relativos, en la forma de ordenamientos de mejor a peor, para selección u otros fines:

- Pruebas de aptitud para ordenar a los candidatos que pretenden entrar a una institución o programa, para seleccionar a quienes habrán de ser admitidos.
- *Rankings* de escuelas en función de los resultados de sus alumnos en algún tipo de prueba o examen, que se publican con la pretensión de ofrecer a las familias información acerca de la calidad de las instituciones.
- Pruebas para seleccionar instituciones para participar en algún programa u otorgarles algún tipo de premio (a las mejores) o apoyo (a las más pobres).

Evaluaciones para tomar decisiones blandas. Buscan servir de base para decisiones y acciones de mejora de aquello que ha sido evaluado. Ejemplos:

- Evaluaciones de alumnos realizadas por su profesor, con el fin de monitorear el proceso de aprendizaje y apoyar a cada uno en función de sus dificultades.
- Evaluaciones de docentes, cuya finalidad es brindarles orientación profesional para desarrollar mejor su labor de enseñanza.
- Evaluaciones de currículo o de proyectos específicos, cuya finalidad es detectar necesidades de cambio y mejora.
- Evaluaciones de logros educativos en gran escala, cuyo propósito es contribuir a identificar las principales debilidades del sistema educativo, y orientar la reflexión y la formulación de políticas educativas.

Evaluaciones para la toma de decisiones institucionales duras. Pretenden servir de base para tomar decisiones que tienen consecuencias importantes y directas para un proyecto o institución. Ejemplos:

- Evaluaciones de programas, cuya finalidad específica es resolver acerca de la continuidad o terminación.
- Las evaluaciones de instituciones o programas con fines de acreditación o certificación podrían caer en esta categoría, pues tienen consecuencias directas e importantes.

Evaluaciones para otorgar incentivos y/o sanciones individuales. Para sustentar la asignación de incentivos o sanciones para la mejora del desempeño de individuos.

- Muchas evaluaciones tienen implícita o explícitamente este propósito, v.gr. aprobar un curso, obtener una buena calificación para acceder a un cargo, desarrollar un mejor trabajo, etc.
- Evaluaciones para asignar incentivos económicos a maestros, v. gr. pagos complementarios en función de: los resultados de los alumnos en pruebas estandarizadas, de sus esfuerzos de capacitación académica, del cumplimiento de obligaciones en materia de puntualidad y asistencia, etc.

Evaluaciones para la rendición de cuentas. En este caso el propósito es aportar elementos para que los responsables de una tarea educativa puedan ser juzgados por quienes tienen derecho a ello, con base en lo que han logrado, haciendo visibles los resultados de su labor, para que rindan cuentas al respecto. Ejemplos:

- Evaluaciones de maestros y escuelas ante familias, administración y sociedad.
- Evaluaciones de la administración educativa ante sociedad y familias.

En una forma más sencilla, pero siempre con base en su propósito, las evaluaciones pueden agruparse en dos categorías: las formativas, que se llevan a cabo a lo largo de un proceso, fundamentalmente buscando aportar elementos de retroalimentación para su mejora; y las sumativas, que se realizan al final de un camino, para sustentar decisiones conclusivas, en términos de aprobación o no aprobación o similares.

Otras clasificaciones son las que distinguen las evaluaciones según criterios técnicos, como pruebas formadas por preguntas de opción múltiple o abiertas; de una forma o matriciales; de pequeña o gran escala; evaluaciones hechas por los docentes en el aula o por instancias externas; de consecuencias fuertes o débiles (alto o bajo impacto), entre otras.

1.3. Implicaciones del propósito para el diseño de una evaluación

La definición de los propósitos de la evaluación tiene importancia decisiva para su diseño, ya que la forma que adopte será distinta según qué preguntas se busque responder, qué consecuencias tendrá, quienes usarán los resultados y para qué. Esta idea quedará más clara considerando los dos ejemplos siguientes.

Pruebas de selección: Las pruebas que se aplican a los aspirantes a entrar a una carrera o institución, implican las siguientes características:

- *Atender la dimensión predictiva de la validez*: dado que se busca identificar a los mejores candidatos del conjunto de los que solicitan admisión en un momento dado, interesa medir el potencial de cada uno para tener éxito en los estudios que quiere iniciar, las cualidades que podrán *predecir* el desempeño de los aspirantes en el futuro.
- *Privilegiar habilidades y no conocimientos*: Las habilidades intelectuales (como competencia lectora, de razonamiento verbal, numérico, etc.) de los aspirantes predicen mejor su rendimiento futuro que su conocimiento de los contenidos puntuales de los planes y programas de los niveles educativos anteriores.
- *No incluir preguntas de dificultad muy baja y muy alta*: Puesto que ni unas ni otras ayudan a distinguir los mejores aspirantes de los menos buenos: en un caso casi todos los aspirantes responden bien; en el otro casi ninguno lo hace.

Pruebas de aprovechamiento: Las que se aplican a los alumnos que terminan un grado o nivel escolar, para ver en qué medida manejan los conocimientos y las habilidades estipuladas por los planes y programas de estudio. Dado este propósito, estas pruebas tienen características opuestas a las de selección:

- *Atender la dimensión concurrente de la validez*: Lo que se quiere medir tiene que ver con algo que ya ocurrió; por ello la validez de las pruebas se detectará valorando la congruencia de los resultados obtenidos en ellas por los alumnos con otras valoraciones de su desempeño, como las calificaciones que les asignaron sus maestros a lo largo del trayecto anterior.
- *Privilegiar contenidos alineados al currículo*: Puesto que se trata de ver si los alumnos que terminan los manejan o no. Si, además, se pretende inferir algo sobre la calidad de los maestros o las escuelas, será importante centrar la atención en los conocimientos y las habilidades cuyo manejo depende más de la escuela, y no de rasgos individuales o de la influencia del contexto familiar, como ocurre con habilidades como las que privilegian las pruebas de selección.
- *Incluir preguntas de todos los grados de dificultad*. Ya que no se busca ordenar a los mejores distinguiéndolos de los menos competentes, sino verificar en qué medida consiguieron todos aprender lo que se pretendía que aprendieran.

1.4. Evaluaciones del aprendizaje que se aplican en México

En seguida se revisan evaluaciones en gran escala del aprendizaje de los alumnos de México, aplicando las ideas anteriores para valorar sus alcances y límites potenciales.

ENLACE

Desde el final del ciclo escolar 2005-2006, la SEP comenzó a aplicar las pruebas conocidas por las siglas ENLACE (Exámenes Nacionales del Logro Académico en Centros Escolares. La decisión de desarrollar estas pruebas se tomó con la idea de que sirvieran sobre todo para que los maestros y los padres de familia de alumnos de primaria y secundaria pudieran tener una información sobre el grado en que cada niño estaba alcanzando un nivel de aprendizaje aceptable en unos cuantos temas importantes de dos áreas clave del currículo correspondiente, para que se pudieran atender las necesidades de apoyo de cada uno oportunamente.

Se trataba, pues, de un propósito eminentemente formativo. La conciencia de que las evaluaciones que hace cada maestro no son agregables, ni utilizan referentes iguales, hacía ver la necesidad de aplicar instrumentos estandarizados en forma censal, o sea a todos los alumnos de ciertos grados de la educación básica. Al mismo tiempo, se tenía conciencia de que el carácter masivo de este tipo de pruebas, en un sistema de las dimensiones del mexicano, hacía imposible el uso de instrumentos que pudieran medir en forma válida y confiable muchos aspectos fundamentales del currículo, que no se prestan para su medición en gran escala con instrumentos integrados únicamente por preguntas de opción múltiple.

Junto con otras consideraciones sobre la necesidad de diseños experimentales y/o longitudinales para valorar la calidad de una escuela y, con mayor razón, de un docente, lo anterior hacía que se tuviera claro que esas evaluaciones censales no podrían ser suficientes ese tipo de propósitos, si bien se sabía que algunos sectores se verían tentados a usar los resultados para ello.

La necesidad de tener resultados por alumno impide el uso de diseños matriciales y hace inevitable el riesgo de que algunos alumnos copien las respuestas de otros, lo que se facilita por las dimensiones del sistema educativo y el carácter masivo de ENLACE, que dificultan el control de su aplicación, por lo que se decidió utilizar procedimientos computarizados para la detección de posibles casos de copia. Estas pruebas, por otra parte, no necesitan ir acompañadas por cuestionarios de contexto, puesto que sus características no las hacen apropiadas para hacer análisis sobre los posibles factores asociados a los resultados de los alumnos, para orientar las políticas respectivas. Por lo mismo es razonable que los resultados se reporten sólo en forma simple, pero para cada alumno y escuela.

Además de cuidar el llamado factor de copia, en el caso de ENLAE es importante también prestar atención a la proporción efectiva de alumnos de un grupo o escuela que presenten efectivamente la prueba, ya que la ausencia de muy pocos niños (dos o tres, o incluso uno) puede sesgar de manera importante los resultados.

Las pruebas ENLACE, por otra parte, no necesitan ir acompañadas por cuestionarios de contexto, puesto que sus características no las hacen apropiadas para hacer análisis sobre los posibles factores asociados a los resultados de los alumnos, para orientar las políticas respectivas. Por lo mismo es razonable que los resultados se reporten sólo en forma relativamente simple, pero para cada alumno y escuela.

Pese a las consideraciones anteriores, el desconocimiento de los alcances y límites de las pruebas ENLACE ha hecho que sus resultados se utilicen para propósitos para los que no son adecuados, en particular para valorar la calidad de una escuela solamente con base en el puntaje promedio obtenido por sus alumnos y, lo que es aún más inadecuado, para valorar el desempeño de los maestros.

No se pretende negar que estos resultados, usados con cautela y junto con otros elementos, pueden ser indicadores interesantes que contribuyan a formarse una idea adecuada de la calidad de escuelas y maestros, pero no en la forma simplista en que está ocurriendo. Al volver de alto impacto pruebas concebidas con propósitos fundamentalmente formativos se corre un riesgo muy grande de que se produzcan consecuencias indeseables, que ya han comenzado a presentarse, como el que los maestros enseñen para la pruebas o que algunas escuelas usen los resultados para campañas mercadotécnicas sin sustento sólido.

EXCALE

En 2004, el INEE comenzó a desarrollar lo que llamó *una nueva generación de pruebas de aprendizaje*, las primeras de las cuales se aplicaron por primera vez en 2005, y se denominaron *Exámenes de la Calidad y el Logro Educativo, EXCALE.*

El propósito de los *EXCALE* es informar sobre el aprendizaje que logran los alumnos del sistema de educación básica como conjunto, no cada uno ni cada escuela, así como identificar los factores más importantes que favorecen o inhiben el aprendizaje, para ofrecer elementos para las decisiones de política educativa.

De este propósito de derivan las características siguientes de las pruebas *EXCALE* que, con excepción de las dos primeras, las distinguen de las pruebas ENLACE:

Alineadas al currículo, porque su propósito es evaluar los aprendizajes estipulados por los planes y programas de estudio oficiales.

Criteriales, porque se diseñan para evaluar el dominio que tienen los estudiantes de un campo del conocimiento en particular, para llegar a juicios de tipo absoluto sobre el grado en que se cumple el currículo, y no simplemente ordenar a los alumnos comparándolos entre sí.

Matriciales, porque pretenden evaluar todos los contenidos curriculares importantes, para lo cual es necesario dividir la prueba en pequeñas porciones, que son respondidas en parte por cada alumno y en conjunto entre todos los alumnos. En 2005 se evaluaron sólo las áreas de matemáticas y español, pero en este último caso se incluyó la expresión escrita, y no sólo la competencia lectora. Luego se han añadido las áreas de ciencias naturales, geografía, historia y educación cívica.

Muestrales, porque no se pretende dar resultados por alumno y escuela, por lo que basta aplicar las pruebas a muestras que, si son bien diseñadas, permiten obtener resultados precisos del conjunto del sistema educativo y de subsistemas como los de las entidades federativas. Este rasgo hace que EXCALE deba tener cuidado para reducir los errores de muestreo, que no importan en el caso de ENLACE; en EXCALE, en cambio, no importan el factor copia ni la proporción efectiva de sustentantes.

Con aplicaciones rigurosamente controladas, a cargo de personal ajeno a la escuela, con base en protocolos de aplicación muy detallados, e incluyendo un monitoreo de la calidad de cada aplicación, a cargo de una instancia externa.

Con preguntas cerradas y abiertas, lo que es posible por la escala mucho menor a la de un censo; las preguntas de respuesta construida son apropiadas para medir los aspectos más complejos del currículo, pero inviables en escala masiva.

Con cuestionarios de contexto, como parte importante de la evaluación, para obtener información de los mismos alumnos, sus maestros, el director de la escuela y, en su caso, de los padres de familia, sobre numerosas variables del contexto de la escuela misma y del hogar.

Con reportes de resultados que incluyen análisis complejos de los obtenidos por los alumnos en las pruebas, así como de los factores del hogar y la escuela asociados con ellos. Para el procesamiento de las respuestas de los alumnos a las pruebas matriciales se utiliza la técnica de valores plausibles; las escalas en que se expresan los resultados, y las de los constructos que se forman a partir de los cuestionarios de contexto, se hacen con base en el modelo de Rasch; para los análisis se utilizan modelos lineales jerárquicos y de ecuaciones estructurales.

Con aplicaciones a un solo grado cada año, conformando un ciclo de cuatro años en los que se aplican pruebas sucesivamente a alumnos de 3° de preescolar, 3° y 6° de primaria y 3° de secundaria. A partir de 2009 se incluirá el 3° de media superior.

PISA

Conviene mencionar que, al igual que EXCALE, las dos pruebas internacionales que en seguida se mencionan tienen como propósito ofrecer un diagnóstico del sistema educativo o subsistemas como tales, para diseño de políticas, y no dar resultados por alumno o por escuela. Por eso tanto las pruebas PISA de la OCDE, como las del Segundo Estudio Regional Comparativo y Explicativo del LLECE tienen rasgos similares a EXCALE. La principal diferencia es que las pruebas de PISA no se construyen con referencia al currículo de ningún país, y las del SERCE lo hacen en relación con los contenidos comunes de los currículos de los países participantes.

El Programa para la Evaluación Internacional de Estudiantes (*Programme for International Student Assessment*, PISA), de la Organización para la Cooperación y el Desarrollo Económico (OCDE), es un esfuerzo de colaboración internacional para monitorear los resultados de los sistemas educativos de los países participantes, miembros y no miembros de la OCDE.

El estudio PISA representa el esfuerzo internacional más completo y riguroso realizado hasta la fecha para evaluar el desempeño de los estudiantes y recabar información sobre los factores –individuales, familiares y escolares—que pueden contribuir a explicar las diferencias de dicho desempeño. En este sentido, ofrece la posibilidad de valorar las fortalezas y debilidades de nuestro sistema educativo en comparación con los de otros países. A partir de sus resultados, PISA se propone ofrecer buenas ideas sobre lo que puede hacerse desde la política educativa para mejorar las oportunidades de aprendizaje de los jóvenes y reducir la desigualdad en los niveles de logro.

El propósito específico de PISA es evaluar las competencias que los jóvenes de 15 a 16 años de edad necesitarán, en su vida adulta, para enfrentar los retos de la sociedad del conocimiento. Las pruebas comprenden tres áreas: lectura, matemáticas, y ciencias y se aplican cada tres años a muestras representativas de escuelas y estudiantes. En cada ciclo se enfatiza un área.

México participa en PISA desde el ciclo 2000, en que se evaluaron 5,276 estudiantes de 183 secundarias y bachilleratos, suficientes para dar resultados representativos a nivel nacional. Para 2003 se amplió la muestra de escuelas y estudiantes, para dar

resultados no sólo a nivel nacional, sino también por entidad federativa: se evaluaron cerca de 30 mil estudiantes de 1,124 escuelas.

Para la aplicación de PISA 2006, se mantuvo la decisión de tener sobremuestra de escuelas (1,140) y estudiantes (30,971) con el propósito de disponer de información representativa por entidad federativa. Además, se decidió que México participara en una opción adicional de evaluación llamada PISA Grado Modal o PISA basada en el grado, cuyos resultados sólo son representativos a nivel nacional. Se definió como grado modal el primer año de bachillerato, debido a que aproximadamente 60% de los estudiantes mexicanos, de acuerdo con las evaluaciones anteriores, se encuentra inscrito en ese grado.

En 2009, la muestra nacional estuvo conformada por poco menos de 53,000 estudiantes de 1,770 escuelas; en esta ocasión, México decidió aplicar las pruebas tanto a jóvenes de 15 años de edad como a estudiantes del último grado de bachillerato.

Desde 2003 el INEE está a cargo de las pruebas PISA en México. Los resultados de las tres aplicaciones que han tenido lugar hasta ahora se difundieron en su momento. En los informes internacionales y los del Instituto pueden verse los de los jóvenes mexicanos, tanto en relación con los estudiantes de los demás países participantes, como por servicio educativo y, desde 2003, por entidad federativa.

En estos informes los resultados se presentan principalmente de dos maneras: a) mediante la puntuación obtenida en promedio por los estudiantes de un país o entidad y b) mediante los porcentajes de alumnos en cada uno de varios niveles de desempeño. Los niveles de desempeño permiten juicios en relación con los niveles deseables, en términos de adecuado o inadecuado, y no de mejor o peor. También se hacen otros análisis para explorar cómo varían los resultados entre países y en cada uno; diferencias de resultados entre tipos de escuelas; y la relación de los resultados con características de los alumnos, sus familias y las escuelas, que se recogen mediante cuestionarios que se aplican junto con las pruebas.

LLECE

En 1994, la Oficina Regional para América Latina y el Caribe de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO-OREALC) promovió la puesta en marcha de un importante proyecto de evaluación, que permitiera comparar el nivel de rendimiento de alumnos de educación primaria de los países de la región.

Considerando que no debería tratarse de un proyecto que se llevara a cabo una sola vez, sino que debería tratarse de una actividad permanente, se formó un grupo con los responsables de evaluación de los ministerios de educación de los países de la región, que tomó el nombre de Laboratorio Latinoamericano de Evaluación de la Calidad Educativa (LLECE). Fue concebido desde su origen como una red de sistemas de medición y evaluación de la calidad educativa, con carácter gubernamental. Se trató de la primera iniciativa de índole estrictamente latinoamericana en el ámbito de los estudios comparativos internacionales del rendimiento académico.

El LLECE puso en marcha el proyecto que, en lo sucesivo, y en referencia al estudio posterior, se designará con el nombre de Primer Estudio Regional Comparativo y Explicativo, PERCE. El estudio implicó el desarrollo y aplicación a muestras nacionales de alumnos de tercero y cuarto grados de primaria, de pruebas de rendimiento en dos áreas: lengua nacional (español o portugués) y matemáticas. El trabajo de campo del PERCE tuvo lugar en 1997, con la participación de trece países de la región, y el informe inicial se publicó en 1998.

En 2004 se inició el Segundo Estudio Regional Comparativo y Explicativo (SERCE). Además de que el estudio busca conocer con mayor precisión lo que saben los estudiantes, también recolecta información que ayuda a identificar los elementos propios de la escuela, del aula y del contexto que puedan contribuir a la explicación de sus rendimientos.

Las áreas de evaluación del SERCE fueron matemáticas, lenguaje (lectura y escritura) y ciencias naturales (optativa) y se aplicó a alumnos de 3° y 6° grados de primaria. La selección de los contenidos se realiza a partir de los currículos de los países participantes. En 2006 participaron 16 países y el estado de Nuevo León. El informe inicial se difundió en junio de 2008.

Además del mayor número de países participantes, del cambio de uno de los grados evaluados, y de la adición opcional del área de ciencias, los instrumentos utilizados en el SERCE fueron diferentes y no equiparables a los del Primer Estudio, por lo que los resultados no se pueden comparar directamente.

2. La investigación sobre evaluación

Por lo que se refiere al segundo aspecto de nuestro tema, el que tiene que ver con las investigaciones que hemos llamado *de segundo orden*. Sobre la evaluación misma, el panorama es diferente, ya que hasta ahora prácticamente no hay trabajos en este sentido, los que parecen muy necesarios, sobre todo en un momento en el que la extensión de las pruebas de aprendizaje en gran escala se está dando en forma notable y, al parecer, no de la manera más adecuada.

Pueden mencionarse varios tipos de investigación sobre la evaluación que podrían hacer aportaciones valiosas al campo:

- En primer lugar pueden mencionarse investigaciones sobre la forma misma en que se llevan a cabo las evaluaciones, para tener elementos sólidos para apreciar objetivamente sus alcances y limitaciones.
- En segundo lugar, investigaciones sobre la forma en que se difunden los resultados, sobre el grado en que llegan a los destinatarios y, sobre todo, el grado en que son interpretados correctamente por ellos. En este apartado se pueden incluir estudios sobre la forma en que los medios de comunicación difunden los resultados de las evaluaciones, los supuestos en que se basan y las orientaciones de política que favorecen, con o sin sustento.
- En tercer lugar, y derivados de lo anterior, estudios sobre el grado en que los resultados son utilizados por los destinatarios para emprender acciones con base en ellos, y sobre la forma en que esto se hace. En este punto se pueden distinguir los usos de los resultados por parte de las autoridades educativas, para el diseño de políticas, pero también el uso por parte de los maestros, para enriquecer su trabajo en aula, e incluso el uso por parte de los padres de familia.

Como se ha dicho, la investigación sistemática sobre los puntos anteriores y otros similares es prácticamente inexistente, y el conocimiento al respecto se limita a impresiones subjetivas.

No cabe duda que es deseable avanzar en este ámbito, lo que parece indispensable para el desarrollo de una sólida *cultura de la evaluación*, entendiendo por esto un conocimiento de las características, alcances y limitaciones de los distintos tipos de evaluación, con base en lo cual tanto las autoridades como los maestros, los padres de familia y la sociedad en general puedan interpretar los resultados de las evaluaciones correctamente, y los puedan utilizar para sustentar decisiones adecuadas, en los respectivos ámbitos de cada uno.

3. Conclusión: reflexiones sobre el potencial de la investigación evaluativa

3.1. Evaluación en grande y pequeña escala

La extensión del uso de pruebas de rendimiento en gran escala en muchos sistemas educativos puede hacerse a partir de dos concepciones muy distintas, que pueden presentarse, de manera simplificada, como sigue:

- En una perspectiva, las pruebas se conciben como la herramienta fundamental para evaluar la calidad de la educación, por encima de otros medios, en especial de las evaluaciones que hacen los maestros mismos, a los que se considera incapaces de hacer evaluaciones suficientemente confiables.
- Las pruebas pueden verse también como un medio entre otros, que puede aportar elementos valiosos sobre algunos aspectos de la calidad educativa, pero siempre incompletos e insuficientes y que, por lo tanto, es necesario que sus resultados se complementen con otros elementos, en particular con los aportados por los maestros, cuyo papel se considera insustituible.

La reciente proliferación de las pruebas en gran escala en los sistemas educativos va frecuentemente acompañada de una difusión de la primera de estas dos posturas, lo que a su vez refleja la extensión de algunas ideas que pueden resumirse como sigue:

- Entre muchas personas interesadas por la educación prevalece un sentimiento de insatisfacción con los niveles de aprendizaje que alcanza la mayoría de los alumnos. Este sentimiento se encuentra en los más diversos países, y se ve reforzado por las interpretaciones usuales de las evaluaciones internacionales.
- De quienes experimentan tal insatisfacción, algunos tienen una visión simplista tanto de las causas de la situación como de la forma de valorarla y corregirla.
    - *Los diagnósticos* no toman en cuenta los diversos factores de las escuelas y su contexto que influyen en la calidad;
    - *Los medios de valorar ésta* se reducen a la aplicación de pruebas de rendimiento, sin distinguir su enfoque y nivel técnico;
    - *Las estrategias para remediar la situación* se reducen a asignar recursos a las escuelas según los resultados de sus alumnos en las pruebas.
- Se observa una proliferación de pruebas con las que se pretende valorar la calidad de las escuelas, aunque los instrumentos no hayan sido diseñados de manera apropiada para tal propósito y aunque, en muchos casos, adolezcan de deficiencias técnicas que ponen en tela de juicio su validez y/o su confiabilidad.

A partir de lo anterior, como se apuntó desde la introducción, se producen dos tipos de reacciones: los maestros se ven impulsados a orientar su trabajo pedagógico en función de los contenidos de las pruebas, y entre los maestros y los estudiosos de la educación se produce un rechazo absoluto de las pruebas, que se consideran inadecuadas para valorar correctamente la calidad educativa.

La postura que sustenta este capítulo se basa en concepciones diferentes en los tres puntos señalados:

- *En cuanto al diagnóstico,* coincide en que, en muchos casos, la calidad de la educación no es satisfactoria, en especial en relación con los países que tienen mejores resultados; antes de calificarla como catastrófica, sin embargo, hay que tener en cuenta muchas cosas, en especial la diferencia fundamental que distingue un sistema educativo que atiende a una minoría, como ocurría hace sólo medio siglo, de otro que atiende a casi toda la población, desde los cuatro años de edad hasta los 14, y cada vez más desde los tres hasta los 17 ó 18.

En lo que se refiere a las causas de la situación, se considera que son complejas y suficientemente conocidas como para que los resultados no resulten sorpresivos para quien tenga una visión amplia del contexto nacional e internacional.

- *En cuanto a los medios de valorar la situación*, su complejidad implica que una evaluación adecuada de la calidad no pueda reducirse a los resultados de los alumnos en pruebas de rendimiento, y menos si éstas no tienen el enfoque y la calidad técnica suficientes. La evaluación que necesita un sistema educativo no puede reducirse a pruebas, aunque debe incluirlas, a condición de que sean de buena calidad técnica y se utilicen de manera parsimoniosa; debe haber indicadores de otras dimensiones de la calidad, estadísticas y acercamientos cualitativos, acordes a la naturaleza de los sujetos y los procesos educativos.
- *Y en cuanto las estrategias de mejora*, dada la complejidad de las causas de la situación prevaleciente deberán ser complejas también; los resultados de tales estrategias sólo podrán observarse en el mediano y largo plazo, como fruto de esfuerzos serios y sostenidos.

3.2. Las pruebas en gran escala como sustitutos del trabajo de los maestros

Nuestra opinión se opone a la postura que reduce y simplifica la evaluación educativa, identificándola con las pruebas en gran escala. Compartimos la idea de que los sistemas educativos necesitan buenas evaluaciones, que ofrezcan a las autoridades, los maestros y la sociedad diagnósticos precisos y confiables, para que puedan fijarse metas ambiciosas y realistas y diseñarse estrategias adecuadas de mejora. Pero subrayamos que debe evitarse un riesgo que acecha a los sistemas de evaluación contemporáneos: el de limitarlos a la aplicación masiva de pruebas de rendimiento, incluso si son de buena calidad técnica, lo que en muchos casos no se asegura.

La extensión de las pruebas en gran escala va acompañada, en muchas ocasiones, de usos inapropiados de los resultados, como los que consisten en hacer ordenamientos simples de escuelas (conocidos como *rankings*) que, supuestamente, reflejarían de manera objetiva la calidad de las escuelas mismas. Con base en ellos las autoridades podrían ofrecer estímulos a las escuelas de mejores resultados, y los padres de familia tendrían una base sólida para decidir a qué escuela enviar a sus hijos. La competencia que se establecería de esta manera entre las escuelas haría mejorar su calidad.

Lo anterior ignora que, al valorar la calidad de las escuelas con base únicamente en los resultados obtenidos por sus alumnos en pruebas estandarizadas, aun suponiendo que éstas sean de buena calidad técnica, se comete un error grave, que pone en cuestión la validez de las inferencias basadas en tales resultados.

Imaginemos dos escuelas. Una es selectiva; admite sólo a los mejores aspirantes, lo que hace que buena parte de sus alumnos provenga de medios favorecidos. Además es exigente, por lo que los alumnos de menor rendimiento no pueden permanecer en ella, sino que la abandonan, sea para ir a otra escuela, sea para dedicarse a otras actividades. La otra escuela acepta a todos los solicitantes de nuevo ingreso, sin selección, lo que hace que una mayoría sea de origen humilde. Además, se esfuerza por mantener hasta el fin del trayecto a todos los aceptados, y lo consigue en gran medida, aunque no todos alcancen plenamente los objetivos de aprendizaje. En una prueba estandarizada, los alumnos de la primera escuela tendrán seguramente resultados superiores, en promedio, a los de los estudiantes de la segunda. ¿Sería adecuado concluir por ello que la primera escuela es mejor que la otra?

Sin más datos, no debería sacarse tal conclusión. Los mejores resultados de los alumnos de la primera escuela pueden deberse simplemente a su extracción social, eventualmente gracias a políticas selectivas, y no a un funcionamiento ordenado o mejores prácticas de enseñanza. Los resultados inferiores de los alumnos de la

segunda escuela, por su parte, podrían ser tales aun en el caso de que la escuela funcione bien, con un trabajo valioso de los docentes, alta participación de los padres y otras prácticas positivas, que explicarían la retención de estudiantes, aunque no se obtengan resultados más altos que los de la otra escuela.

Las estrategias de mejora basadas en la asignación de estímulos económicos o el establecimiento de una competencia entre las escuelas, con base únicamente en los resultados de pruebas, parten de una transferencia poco sustentada de los principios de la economía, y no tienen en cuenta las peculiaridades de la oferta y la demanda educativas, que no siguen necesariamente la lógica del mercado. Dichas estrategias, además, ignoran las dificultades reales y considerables que representan las desigualdades sociales, para el propósito indiscutible de que los alumnos de todas las escuelas de un país consigan resultados similares.

En otras palabras, las estrategias simplistas de mejora parten de un supuesto falso: que hacer buena educación en cualquier contexto es fácil: *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 de rendimiento, y asocie consecuencias fuertes a las pruebas, en la forma de premios cuando los resultados suben y sanciones cuando no ocurra así*. (Hamilton, Stecher y Klein, 2002)

Algunos usos de resultados de pruebas en gran escala pueden tener, además, serias consecuencias negativas para la calidad educativa misma. La asignación de estímulos económicos con base en ellos y los ordenamientos simples de escuelas hacen que las pruebas se vuelvan *de alto impacto*, lo que propicia que se corrompan, al aparecer prácticas negativas como la preparación de alumnos para la prueba, la subordinación del currículo a las evaluaciones, o la alteración de resultados mediante estrategias más abiertamente deshonestas. Un destacado especialista americano se refiere a estas consecuencias indeseables en los términos siguientes:

> *Por la errónea utilización de pruebas de rendimiento estandarizadas tradicionales para evaluar la calidad de las escuelas hay cosas realmente terribles que están ocurriendo en las escuelas de nuestros niños en estos días.*
>
> *Una es que aspectos importantes del currículo se están haciendo a un lado, porque no son medidos por las pruebas. Otra es que los niños están siendo entrenados sin descanso para que dominen el contenido de esas pruebas de alto impacto y, en consecuencia, están comenzando a odiar la escuela. Y una más es que, en muchos casos, los maestros se dedican a preparar a sus alumnos para las pruebas, lo que se parece mucho a hacer trampa, porque están inflando las puntuaciones de los alumnos sin elevar su competencia en los aspectos que se supone miden las pruebas…* (Popham, 2001)

3.3. Las pruebas en gran escala como apoyo al trabajo de los maestros

La segunda postura, que compartimos, ve a las pruebas en gran escala como un medio útil para complementar el trabajo de los maestros, pero no como un sustituto del mismo. Partimos de la idea de que el trabajo de un buen docente es insustituible tanto para conseguir que los alumnos alcancen un alto nivel de competencia en cuanto a los conocimientos y habilidades que necesitarán para una vida plena en el mundo de hoy, como para valorar el grado en que tal cosa ocurre, o sea para evaluar.

La tarea de evaluar el grado en que un alumno ha desarrollado los conocimientos y habilidades previstos al final de un ciclo escolar no es sencilla, si se quiere cubrir bien las áreas del currículo y los temas de cada una. Para conocer el avance del alumno la tarea se complica, ya que se deberá evaluar al inicio y al fin del ciclo escolar. Si se quiere conocer con más detalle el avance de un alumno en intervalos menores (mensual o incluso semanalmente), la tarea se vuelve mucho más compleja.

Si en lugar de un alumno se trata de dos o tres decenas, y además se quiere tener información sobre las circunstancias personales, familiares y sociales de cada uno de ellos, para tenerla en cuenta en el momento de tomar decisiones importantes para el futuro de cada uno de ellos, la tarea se antoja difícil.

Eso es lo que se espera de los maestros, y es crucial para que el trabajo educativo tenga buenos resultados: para retroalimentar su propio trabajo docente, así como el esfuerzo de los alumnos mismos, es fundamental que el maestro conozca con precisión el avance de cada uno de sus alumnos. Por ello la calidad de un sistema educativo se basa en última instancia en el profesionalismo de sus maestros que, además de dominar los contenidos a enseñar y los métodos pedagógicos necesarios para ello, deben también ser capaces de manejar técnicas de evaluación apropiadas para el trabajo en el aula, que les proporcionen la información necesaria para retroalimentar su propio trabajo y el de los alumnos.

Sabemos que muchos maestros del sistema educativo mexicano, posiblemente una gran mayoría de ellos, no evalúan el avance de sus alumnos en la forma profunda y precisa a la que se alude en los párrafos anteriores, y sabemos que las deficiencias de la formación de los maestros, las circunstancias desfavorables en que se desarrolla el trabajo de muchos de ellos, entre otros elementos, explica la situación.

La experiencia muestra también sin embargo, que si bien no es sencillo que los maestros hagan buenas evaluaciones, sí es posible, y que la evaluación que hace un buen maestro del avance de sus alumnos tiene niveles de validez y confiabilidad suficientes para sustentar las decisiones educativas más delicadas.

La pregunta siguiente es: ¿Podrá evaluarse el aprendizaje con validez y confiabilidad comparables con pruebas de gran escala? La respuesta es clara: con la finura que puede alcanzar la evaluación a cargo del maestro no es posible con las metodologías de evaluación en gran escala disponibles; probablemente no llegue a serlo nunca con la profundidad de que se trata. Las evaluaciones de aprendizaje en gran escala pueden, en cambio, dar información de buena calidad sobre conjuntos grandes de alumnos, en ciertas áreas del currículo y con intervalos de tiempo amplios.

La mayoría de las evaluaciones en gran escala que se aplican en la actualidad utilizan pruebas estandarizadas compuestas por preguntas de opción múltiple. Se usan también, aunque con frecuencia mucho menor, instrumentos menos estructurados: preguntas de respuesta construida, ejercicios con problemas o situaciones reales, así como evaluaciones orales y observaciones de las ejecuciones de los evaluados. El uso de preguntas de opción múltiple en evaluaciones en gran escala no es accidental: aunque tienen limitaciones para la evaluación de niveles cognitivos complejos, su viabilidad para aplicaciones a números grandes de sujetos es mucho mayor que el de acercamientos menos estructurados. El uso de acercamientos alternativos aún con pocos cientos de alumnos implicaría muchas horas de trabajo de un gran número de calificadores especializados, lo que las hace inviables para su uso en gran escala.

El desarrollo de pruebas de buena calidad basadas en preguntas de opción múltiple es también laborioso, pero una vez desarrolladas, pueden aplicarse a miles de sujetos en forma controlada, lo que se traduce en costos unitarios bajos. Este tipo de evaluaciones tiene, sin embargo, limitaciones que deben tenerse en cuenta para entender para qué tipo de propósitos pueden usarse en forma adecuada.

Ninguna decisión importante sobre alumnos o escuelas en lo individual debería basarse únicamente en los resultados de unas pruebas estandarizadas; estos últimos son esenciales para la evaluación de los grandes conjuntos que son los sistemas educativos, ya que las evaluaciones de los docentes no pueden agregarse por su inevitable carácter contextual; los resultados de pruebas estandarizadas son valiosos también como complementos de la evaluación individualizada a cargo de maestros,

directores y supervisores, en especial para que haya referentes sólidos sobre los niveles promedio alcanzados en escala macro, para comparar el nivel de cada alumno.

Las posturas simplistas sobre la evaluación en gran escala ignoran esta complejidad y sobreestiman las posibilidades de los instrumentos usuales; pierden de vista que el maestro y los padres son y serán piezas clave en la búsqueda de mejora educativa. Por ello consideramos que la postura adecuada de las dos mencionadas es la que concibe el papel de las pruebas en gran escala como complemento del trabajo de los maestros. Así y sólo así la evaluación será una herramienta que contribuya efectivamente al mejoramiento de las escuelas, y que sea apreciada como tal por los maestros y por todas las personas preocupadas por la educación.

3.4. La *revolución cognitiva* y la evaluación en pequeña y gran escala

La Teoría Clásica de los Tests y las pruebas de rendimiento en gran escala de diseño tradicional se desarrollaron, como se ha dicho, durante la primera mitad del siglo XX; ambas estuvieron marcadas por las concepciones psicológicas y pedagógicas de la época, entre las que destacaban corrientes como el conductismo de Skinner. Los avances de las nuevas concepciones psicométricas, de mediados del siglo pasado en adelante, se dieron a su vez en forma paralela a la llamada *revolución cognitiva*, de la que se derivan también las corrientes pedagógicas que se engloban bajo la etiqueta demasiado trillada del *constructivismo*.

Estos desarrollos coinciden en rechazar el planteamiento conductista que reduce el campo de estudio de la psicología a los fenómenos más directamente observables, para intentar *abrir la caja negra de la mente*, explorando los procesos que tienen lugar en su interior, con técnicas como las de *pensar en voz alta*. La revolución cognitiva, dice Lorrie Shepard, fue:

> *…una rebelión contra la psicología de las diferencias individuales y el conductismo, una de cuyas premisas básicas era el centrar la atención en la adquisición de competencias gracias al refuerzo de conductas observables y no en tratar de explicar los procesos mentales subyacentes*. (2006: 627)

En la medida en que se identifican y exploran los procesos mentales –y los avances de las ciencias cognitivas durante las últimas cinco décadas muestran que lo es en un grado considerable— se abren horizontes vastos y atractivos tanto para la pedagogía como para las metodologías de evaluación del aprendizaje.

La creciente literatura relativa a la llamada *evaluación en aula* (*classroom assessment*) muestra la riqueza potencial de estos desarrollos. (Cfr. Allal y Mottier, 2005; Black y Dylan, 2005; Köller, 2005; además de Shepard, 2006)

Según esta última autora, los impulsores iniciales del desarrollo de las pruebas estandarizadas, estaban convencidos de que las escuelas americanas tenían problemas de calidad serios, y se propusieron desarrollar instrumentos que permitieran comparar los resultados obtenidos por los alumnos de diferentes escuelas. Shepard cita la opinión de Thorndike, en el sentido de que las nuevas pruebas serían *un remedio para la escandalosa falta de confiabilidad de los exámenes aplicados por los maestros, demostrada en varios estudios previos*. (Shepard, 2006: 623)

Aunque ya en 1923 B. D. Word se quejaba de que las pruebas estandarizadas medían *sólo hechos aislados y piezas de información, en lugar de capacidad de razonamiento, abilidad organizadora, etc.* los nuevos instrumentos se ganaron pronto un lugar en los medios educativos americanos, por sus claras ventajas prácticas. Algunos de los más destacados impulsores de estos instrumentos, como Ralph Tyler, subrayaron siempre, sin embargo, la necesidad de verlos *no como un proceso separado de la enseñanza, sino como parte integral de ésta*.

Pese a ello, la tendencia dominante fue la de considerar las pruebas en gran escala como la forma preferida de evaluación, en tanto que la que realizan diariamente los maestros en las aulas se veía como una forma secundaria, que debería subordinarse a la primera, cuyos principios metodológicos debía imitar.

El contenido de los textos sobre evaluación utilizados en las instituciones formadoras de maestros así lo muestra: según estas obras, las evaluaciones que deberían aplicar en el aula los maestros debían ser réplicas de las evaluaciones en gran escala, por lo que los maestros debían aprender a elaborar preguntas estructuradas y a analizar los resultados de instrumentos formados con ellas estadísticamente, cuidando la validez y la confiabilidad en la misma forma en que debe hacerse en gran escala. (Cfr. Shepard, 2005: 623-625)

Sólo en una época relativamente reciente, *los especialistas en medición comenzaron a prestar atención al contexto del aula para entender mejor las necesidades de los maestros en lo relativo a la preparación para llevar a cabo evaluaciones*.

Siempre según Shepard, Dorr y Bremme:

> *…concluyeron que los maestros piensan con base en razonamientos prácticos y actúan como clínicos, orientando sus actividades evaluadoras a sus tareas cotidianas, como decidir qué enseñar, y cómo hacerlo con alumnos de diversos niveles de desempeño; monitorear el progreso de los alumnos, para saber cómo ajustar la enseñanza en consecuencia; y asignar calificaciones a sus alumnos con base en su desempeño*. (2005: 625-626)

Shepard afirma además que:

> *…la evaluación no puede promover el aprendizaje si se basa en tareas o preguntas que distraen la atención de los objetivos reales de la enseñanza. Históricamente, las pruebas tradicionales muchas veces orientaban la instrucción en una dirección equivocada, si centraban la atención en lo que es más fácil de medir, en vez de hacerlo en lo que es más importante de aprender*.
>
> *…en forma separada de la literatura especializada en medición, los expertos en contenidos curriculares comenzaron también a desarrollar alternativas a las pruebas estandarizadas para su uso en evaluaciones en el contexto del aula, movidos tanto por el rechazo de los efectos de las pruebas utilizadas para rendición de cuentas, como por los profundos cambios en las concepciones del aprendizaje y del manejo adecuado de los contenidos*. (2006: 626)

Según Shepard, en 1989 Silver y Kilpatrick sostenían que,

> *…más allá de la práctica prevaleciente según la cual los maestros desarrollan sus propias pruebas para que se parezcan, tanto en forma como en contenido, a las pruebas de opción múltiple externas, debería hacerse un serio esfuerzo para prepararlos más bien para que puedan conducir lecciones de solución de problemas, y para evaluar la habilidad y las disposiciones de sus alumnos al respecto en el marco de esas lecciones*.

Y la multicitada autora concluye diciendo:

> *El nuevo modelo de evaluación formativa aspira a hacer de la evaluación una parte integral de la enseñanza… La diferencia fundamental consiste en que las nuevas estrategias se basan en un modelo de enseñanza y aprendizaje muy diferente, y no se basan en instrumentos estandarizados desarrollados fuera del aula*. (2006: 627)

3.5. La evaluación de otros aspectos del sistema educativo

La convicción de que todo instrumento de evaluación es necesariamente limitado, apropiado para ciertos propósitos y circunstancias e inapropiado para otros y otras, lleva a una conclusión fundamental: que para evaluar integralmente una realidad compleja deben combinarse varios acercamientos, cuyos alcances y limitaciones se complementen de forma que en conjunto den una visión más amplia que permita formular juicios de valor que sustenten adecuadamente las decisiones. Este principio se aplica tanto a la evaluación de alumnos, como a la de maestros y la de escuelas o planteles como tales, como se desarrolla brevemente en los incisos siguientes.

Valorar integralmente el desempeño de un solo estudiante implica considerar todas las áreas del currículo; aspectos cognitivos y no cognitivos, simples y complejos; al final del ciclo, al inicio y a lo largo del mismo; los factores que favorecen u obstaculizan el avance del alumno, etc. Una valoración así sólo puede hacerla bien un buen maestro.

La evaluación que hace el maestro, aunque es la más completa, tiene una limitación derivada de su misma naturaleza integral y contextualizada: que no puede agregarse en gran escala. Por ello las pruebas estandarizadas, como se ha dicho ya, aportan algo que los maestros no pueden: la posibilidad de comparar grandes grupos en forma confiable, pero no pueden sustituir al maestro en la evaluación integral de los alumnos. Las pruebas en gran escala deben complementar las evaluaciones del maestro, no sustituirlas. La imposibilidad de atender todos los propósitos con un solo instrumento lleva al desarrollo de pruebas distintas, según el propósito que se persiga. Una posibilidad interesante es la de combinar, por una parte, pruebas aplicadas a muestras de alumnos, con un diseño que no permita dar resultados de cada uno en lo individual, ni de cada escuela, pero que cubran de manera muy completa los contenidos curriculares, gracias a un diseño matricial; por otra parte, pruebas más cortas, de menor cobertura curricular, pero que se apliquen en forma censal a todos los alumnos de ciertos grados, permitiendo dar resultados individuales y por escuela.

De manera análoga, la valoración integral de la calidad de un maestro sólo podrá ser el resultado del contacto amplio de un profesional competente con el evaluando, que permita la recolección de evidencias sólidas del desempeño en múltiples aspectos. Esa persona es, normalmente, el director de escuela. Aunque tampoco es sencillo, es más factible que las evaluaciones de docentes hechas por los directores sean comparables, si se utilizan metodologías e instrumentos similares, y se emplean técnicas de *moderación*.

Los resultados de los alumnos en una prueba estandarizada pueden ser útiles para la evaluación de su maestro, pero por sí solos no son suficientes, dados los numerosos factores que inciden en el rendimiento escolar. La aplicación de pruebas a los docentes mismos, si se cuida la validez, puede aportar elementos valiosos y comparables, aunque no podrá captar dimensiones básicas de la práctica docente, que sólo pueden apreciarse mediante acercamientos de observación, difícilmente utilizables en gran escala.

Por lo anterior la evaluación de maestros no puede basarse sólo en la aplicación de instrumentos estandarizados, ni sólo en los resultados de los alumnos, sin tener en cuenta otros aspectos que implican el uso de acercamientos más intensivos, como los portafolios de evidencias, así como la intervención de directores y supervisores. En los niveles superiores del sistema educativo, la edad de los alumnos permite utilizar encuestas que recojan su opinión sobre los docentes, así como la evaluación de pares, en especial para valorar cualitativamente la producción académica.

En forma similar, la valoración integral de la calidad de una escuela sólo puede ser el resultado de un cuidadoso y amplio contacto con el plantel por parte de un profesional capaz de observar, registrar, sistematizar y valorar evidencias de las múltiples facetas

de la calidad educativa en ese nivel. Los sistemas educativos suelen contar para ello con la figura del inspector o supervisor.

Como se ha apuntado, en cambio, reducir la evaluación de la calidad de las escuelas a ordenamientos basados en los resultados de pruebas, sin tener en cuenta otros elementos y sin asegurar la confiabilidad y validez de los resultados es inapropiado y llevará a distorsiones perniciosas para el funcionamiento de las escuelas.

Conviene destacar la importancia que tiene la evaluación de las escuelas como tales; cada una es un organismo con características propias. En las de organización completa, máxime si tienen varios grupos en cada grado, el papel del equipo docente y del director a su cabeza, es crucial para un funcionamiento que se traduzca en mejores niveles de calidad. Por ello la evaluación de las escuelas reviste una importancia fundamental. De hecho la calidad de la evaluación de maestros y alumnos depende fundamentalmente del funcionamiento de la escuela.

3.6. El cambio de paradigma en evaluación

Se ha mencionado ya el parentesco que hay entre las concepciones que sustentan el desarrollo de las pruebas estandarizadas de tipo tradicional y las ideas psicológicas dominantes a mediados del siglo XX, así como las implicaciones para la evaluación de las concepciones contemporáneas, derivadas de la revolución cognitiva. Se han apuntado también los riesgos que entrañan las evaluaciones deficientes y las perspectivas que abren las nuevas concepciones. La autora ampliamente citada antes dice en este sentido:

> *…un extenso cuerpo de literatura ha documentado los efectos negativos sobre la enseñanza y el aprendizaje producidos, ante todo, por el efecto desorientador de enseñar en función de pruebas de formatos limitados y representaciones deficientes de los objetivos de aprendizaje más significativos*. (Shepard, 2006: 639)

Y cita a los autores de una obra reciente (Pellegrino et al. 2001), quienes contemplan

> *…un futuro con un sistema de evaluación más coherente y equilibrado, en el que la evaluación formativa en el aula reciba la misma atención que las pruebas externas de alto impacto, y en el que la evaluación que se hace en el interior del aula y la externa se articulen coherentemente a un mismo modelo de aprendizaje que las soporte a las dos*. (Shepard, 2006: 639)

En las conclusiones de su trabajo, Shepard afirma:

> *La nueva visión de la evaluación en el aula presentada en este capítulo es dramáticamente diferente de la imagen de aplicación en el aula de pruebas estandarizadas, de los volúmenes previos de Educational Measurement y los libros de texto sobre medición. Esta nueva visión centra la atención en una conceptualización mucho más rica del aprendizaje de los alumnos, en el contexto de actividades significativas, y subraya el uso formativo de las evaluaciones, para mejorar el aprendizaje*. (2006: 640)

No hay que olvidar otras dimensiones de la evaluación de la calidad educativa, en especial la evaluación de los docentes mismos, y la de los centros escolares como tales, de las que se ha hablado poco en este trabajo. Lo que se ha dicho es que la evaluación de docentes y la de centros escolares no pueden basarse única ni principalmente en los resultados de evaluaciones en gran escala del aprendizaje, sino que implican acercamientos finos, que consigan entender los procesos correspondientes, en forma análoga a lo que busca el nuevo paradigma de la evaluación en aula respecto del aprendizaje de los alumnos.

El trabajo de los directores y los equipos docentes de centro, así como el de los supervisores, es insustituible para estas evaluaciones. El cambio de paradigma de la evaluación no estará completo si no incluye también estas dimensiones, además de la evaluación de aprendizajes.

El desarrollo de un sistema de evaluación acorde con estas líneas generales sería, a nuestro juicio, un componente esencial de la transformación cualitativa que, sin duda necesitan nuestros sistemas educativos.

La forma ideal de analizar la evolución de los niveles de aprendizaje que alcanzan los alumnos implica la realización de estudios longitudinales, en el sentido estricto de este término: trabajos en los que se sigue individualmente a lo largo de los ciclos escolares a un grupo de estudiantes, con lo que se pueden controlar rigurosamente no sólo las variables del entorno y la escuela que influyen en el rendimiento, sino también las características personales de los sujetos.

Para poder comparar adecuadamente resultados obtenidos en diferentes momentos del trayecto escolar de los alumnos, este tipo de estudios necesita que las mediciones del aprendizaje se hagan con instrumentos cuyos resultados puedan ponerse en una misma escala, o sea que estén *equiparados verticalmente*. Si además se utiliza un diseño experimental o, al menos, se controla estadísticamente el efecto de variables importantes del entorno, se podrá llegar a conclusiones sólidas de tipo causal, sobre el efecto real de la escuela en los resultados de los alumnos, en contraposición al impacto del entorno familiar y social.

Estos requisitos, sin embargo, son difíciles de cumplir en la práctica, por lo que tales estudios son escasos; en México no contamos todavía con trabajos de esas características.

3.7. Evaluación y medición

Para tener buenas evaluaciones no basta partir de un concepto rico de calidad educativa. Es necesario tener también un concepto adecuado de evaluación, lo que no siempre se da, ya que suele haber concepciones limitadas al respecto. Si se aplican pruebas para medir el aprendizaje, por ejemplo, suele pensarse que se ha *evaluado*, cuando en realidad sólo se ha *medido*. *Evaluar* exige algo más: comparar el resultado de la medición con un punto de referencia que establezca lo que se debería saber, para llegar a un juicio sobre lo adecuado o inadecuado del aprendizaje.

El concepto de evaluación que se propone es más amplio. Establece que *la evaluación es el juicio de valor que resulta de contrastar el resultado de la medición de una realidad empírica con un parámetro normativo previamente definido*.

Una buena evaluación no debe caracterizarse solamente por cualidades técnicas como confiabilidad y validez. **Estas dos características son indispensables en toda buena medición, y como evaluar implica medir, la evaluación también las requiere.** Pero como evaluar va más allá de medir, una buena evaluación implica también otras características:

- *Carácter comprensivo de la conceptualización que la sustente*, que deberá atender todas las dimensiones de la calidad educativa, y no sólo algunas, como el nivel de aprendizaje. Es por ello que un sistema de evaluación no puede limitarse a pruebas de rendimiento.
- *Alto nivel técnico de las mediciones en que se base*, para garantizar su confiabilidad y validez, mediante diseños sólidos; el uso de enfoques diversos y complementarios en modelos e instrumentos; la selección cuidadosa de muestras representativas; rigurosos procesos de recolección de datos; y análisis cuidadosos de los datos obtenidos.

- *Pertinencia de los referentes que se definan como parámetros* para contrastar con ellos los resultados de la medición, de manera que las comparaciones tengan sentido. Los referentes se definen normativamente, no se derivan de los datos empíricos. Puede ser *óptimos*: ideales con que se compara una situación; *promedios* de los individuos que se evalúa; y *mínimos*, con los menores valores aceptables. Cada uno arroja cierta luz sobre lo evaluado y ninguno es suficiente. Convendrá usar los tres tipos de parámetro, para una mejor apreciación.
- *Mesura de los juicios de valor* derivados de contrastar mediciones y parámetros, que evitarán excesos triunfalistas o derrotistas y tendrán siempre en cuenta el valor de equidad, considerando el contexto de los alumnos y las escuelas.
- *Amplitud, oportunidad y transparencia de la difusión de resultados*, que deberá llegar a los sectores involucrados en versiones adecuadas a las características de cada uno.

REFERENCIAS

ALLAL, LINDA y LUCIE MOTTIER LOPEZ (2005). Formative Assessment of Learning: A Review of Publications in French. En CERI, 2005: 241-264.

BLACK, PAUL, y WILIAM DYLAN (2005). Changing Teaching through Formative Assessment: Research and Practice. En CERI, 2005: 223-240.

CENTRE FOR EDUCATIONAL RESEARCH AND INNOVATION (2005). *Formative Assessment. Improving Learning in Secondary Classrooms*. Paris. OCDE.

DE LANSHEERE, GILBERT (1986). *La récherche en éducation dans le monde*. Paris. Presses Universitares de France. Traducción al español, con un capítulo sobre la investigación educativa en México y América Latina por Felipe Martínez Rizo, México, Fondo de Cultura Económica, 1996.

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

INEE (2003) *La calidad de la educación básica en México. Primer Informe Anual 2003*. México. Instituto Nacional para la Evaluación de la Educación.

KÖLLER, OLAF (2005). Formative Assessment in Classrooms: A Review of the Empirical German Literature. En CERI, 2005: 265-279.

MARTÍNEZ ARIAS, ROSARIO (1995). *Psicometría: teoría de los tests psicológicos y educativos*. Madrid. Síntesis.

MARTÍNEZ RIZO, FELIPE (2005). “Sobre la difusión de resultados por escuela”. *Cuadernos de Investigación*, N° 18. México, INEE.

POPHAM, JAMES W. (2001) *Frontline: testing our schools: interviews: James Popham* http://pbs.org/html

RAVELA, PEDRO

SHEPARD, LORRIE A. (2006). Classroom Assessment. En Robert L. Brennan, Ed. *Educational Measurement*. 4th Ed. Westport, CT. Praeger. Págs. 623-646.