Martínez Rizo, F. (2009). Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado. *Revista Electrónica de Investigación Educativa, 11* (2). Consultado el día de mes de año en: http://redie.uabc.mx/vol11no2/contenido-mtzrizo2.html

Revista Electrónica de Investigación Educativa

Vol. 11, No. 2, 2009

Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado

Classroom Evaluation for Training, and Large-Scale Evaluation: Toward a More Balanced System

Felipe Martínez Rizo
fmrizo@prodigy.net.mx
Programa de Doctorado Interinstitucional en Educación
Universidad Autónoma de Aguascalientes

San Cosme 108, 20010
Aguascalientes, Aguascalientes, México

(Recibido: 21 de mayo de 2009; aceptado para su publicación: 21 de julio de 2009)

Resumen

Ante la proliferación de pruebas estandarizadas a gran escala que ha tenido lugar en México en los últimos años, el artículo constituye una revisión de la literatura internacional, para reflexionar sobre las posibles consecuencias de ese fenómeno y explorar los avances de enfoques alternativos de evaluación. Se revisa también el desarrollo de las concepciones relativas a la evaluación en aula con propósitos formativos, y se sintetizan las ideas actuales al respecto. Se subrayan la importancia que tales acercamientos pueden tener, por lo que se refiere a la mejora de la calidad educativa. Para concluir, se sostiene que es necesario avanzar en dirección de sistemas de evaluación que combinen de manera más equilibrada la evaluación a gran escala y la evaluación en aula.

Palabras clave: Evaluación educativa, pruebas estandarizadas, evaluación formativa.

Abstract

Given the large-scale proliferation of standardized tests that has occurred in Mexico in recent years, this article constitutes a review of the international literature on the subject, for the purpose of reflecting on the possible consequences of this phenomenon and exploring the progress of alternative assessment approaches. It also reviews the development of concepts relating to evaluation in the classroom for training purposes, and summarizes current thinking about this. It emphasizes the importance that such approaches may have, as regarding the improvement of educational quality. In conclusion, it argues that it is necessary to move toward evaluation systems that combine large-scale assessment and classroom assessment in a more balanced fashion.

Key words: Educational assessment, standardized tests, formative evaluation.

Introducción: Evaluación del aprendizaje y pruebas estandarizadas

La evaluación del aprendizaje tiene antecedentes antiguos. En China comenzaron a aplicarse pruebas a grandes números de personas más de 1,000 años A.C, (Oakes y Lipton, 2007). Mucho después, en el siglo XVI de nuestra era, los liceos jesuitas iniciaron una tradición que, en el XIX, llevó a exámenes *tipo ensayo*, como el *abitur* alemán o el *baccalaureat* francés.

En las escuelas elementales la evaluación se sistematizó más tarde, ya que los sistemas educativos en esos niveles sólo se consolidaron después de que la revolución industrial y la ilustración hicieron que se considerara necesario que todos los futuros ciudadanos supieran al menos leer y escribir. Antes la enseñanza de primeras letras estaba a cargo de preceptores privados en hogares acomodados, o se daba en pequeñas escuelas parroquiales o gremiales. El número de alumnos era reducido y no existía la noción de grado. La evaluación no implicaba el uso de procedimientos sistemáticos; bastaba el juicio del maestro, que no necesitaba usar instrumentos especiales; era suficiente la observación cotidiana que el docente tenía del progreso de cada uno de sus estudiantes.

Cuando los niños que aprendían a leer y escribir eran una minoría, su nivel era también menos heterogéneo que hoy, y los estándares de calidad que un maestro utilizaban implícitamente al formular juicios de evaluación eran relativamente simples. Al generalizarse el acceso a la educación el alumnado se volvió también más heterogéneo, y fue más difícil mantener estándares de calidad comparables.

En Estados Unidos se desarrolló, desde el siglo XIX, un sistema de educación de cobertura masiva, no sólo en educación básica, sino también en media y superior. Por ello, no sorprende que en ese país surgieran tempranamente evaluaciones a gran escala, con la aplicación de pruebas de historia a más de 500 escolares de Boston, en 1845. En 1895 Rice aplicó pruebas de ortografía a 16,000 alumnos, y en 1897, de aritmética a 13,000 estudiantes, y de lectura a 8,300 (De Landsheere, 1986/1996).

En 1890 J. McKeen Cattell, con su artículo *Mental tests and measurements,* inventó la palabra *test* y publicó ese texto fundacional. Binet desarrolló unas pruebas de inteligencia, que luego fueron adaptadas por Terman en Stanford en 1916, y se extendieron con el Army Test en 1917 (De Landsheere, 1996).

Gracias al desarrollo de la psicometría, el College Board –organismo especializado que fue creado en 1900 para elaborar pruebas de ingreso comunes para un grupo de universidades de la costa este de los Estados Unidos– estuvo en condiciones, en 1925, de desarrollar pruebas de aptitud (en contraposición a las de conocimientos), que iban más allá de la memorización de datos aislados y se acercaban a la evaluación de habilidades intelectuales básicas. Desde la década de 1920 en la Universidad de Princeton se hicieron trabajos en este campo, y en 1948, la oficina que se dedicaba a la elaboración de tests se separó de la universidad, constituyéndose el Educational Testing Service (De Landsheere, 1986).

En la segunda mitad del siglo XX el American College Testing (ACT) y la Universidad de Iowa desarrollaron también pruebas importantes. Hasta esas fechas, sin embargo, casi únicamente en el ámbito anglosajón hubo avances similares, al grado de que la psicometría se llegó a considerar una disciplina estadounidense. Esta situación llegó a ser tan marcada que en 1931, al escuchar que los participantes en un congreso se referían a la psicometría como estadounidense, E. L. Thorndike protestó diciendo que “por el bien de la ciencia y por el nuestro, sería preferible que las pruebas estandarizadas no fueran denominadas 'exámenes estadounidenses’”(Joncich, como se cita en De Landsheere, 1996).

I. Predominio y extensión de la evaluación a gran escala

Los pioneros de las pruebas estandarizadas estaban convencidos de que las escuelas tenían serios problemas de calidad, y consideraban que las evaluaciones de los maestros tenían deficiencias graves. Por ello, buscaron elaborar instrumentos que permitieran comparar los niveles de rendimiento de alumnos de diferentes escuelas. Thorndike pensaba que las pruebas remediarían la *escandalosa falta de confiabilidad de los exámenes aplicados por los maestros* (Shepard, 2006, p. 623).

La ventaja de la comparabilidad de los resultados que ofrecían los nuevos instrumentos era atractiva, pero sus limitaciones fueron advertidas desde fechas tempranas. En 1923, B. D. Wood se quejaba de que las pruebas estandarizadas medían sólo *hechos aislados y piezas de información,* en lugar de *capacidad de razonamiento, habilidad organizadora*, etc. Ralph Tyler, subrayó también desde los primeros años la necesidad de verlas no como un proceso separado de la enseñanza, sino como parte integral de ésta (Shepard, 2006).

El contenido de los textos sobre evaluación utilizados en las instituciones formadoras de maestros, muestra que prevalecía la idea de que las evaluaciones

que los maestros debían aplicar en el aula debían ser réplicas de las evaluaciones a gran escala. Por lo tanto, los maestros debían aprender a elaborar preguntas estructuradas y a analizar los resultados de instrumentos formados estadísticamente con ellas. Además, debían cuidar la validez y confiabilidad de tales instrumentos, en la misma forma en que debe hacerse a gran escala. (Shepard, 2005).

Varios acontecimientos contribuyeron a generar un clima de preocupación sobre la calidad de la educación que las escuelas norteamericanas ofrecían a los niños, al comenzar la segunda mitad del siglo XX. Entre dichos acontecimientos pueden mencionarse: el impacto del lanzamiento del Sputnik por la entonces Unión de Repúblicas Socialistas Soviéticas (URSS), en 1957; el Informe Coleman, en 1966, y la tendencia a la baja de los resultados promedio obtenidos año tras año por los aspirantes a ingresar a la educación superior en el Scholastic Aptitude Test (SAT).

El National Defense Education Act, de 1958, muestra el lugar de la educación en la lectura del lanzamiento del Sputnik, en el contexto de la guerra fría (Mathison y Ross, 2008). En los años siguientes, las legislaturas de California, Florida y Oregon establecieron la obligación de evaluar a los alumnos mediante pruebas construidas en relación con estándares mínimos de desempeño (*minimum competency testing*), como parte importante de sus estrategias de mejora.

Para 1982, 42 de los 50 estados norteamericanos tenían programas obligatorios de esa naturaleza. Al generalizarse, las *pruebas de competencias mínimas* muchas veces se hicieron de manera deficiente, por lo que su impacto se redujo y las expectativas depositadas en ellos no se cumplieron (Baker y Choppin, 1990).

La publicación del informe *A nation at risk*, en 1983, mostró la continuidad de la preocupación norteamericana por la calidad educativa, en una perspectiva de seguridad nacional. Con él inició el movimiento de *estándares educativos*, que se manifestó con fuerza durante la década de 1990 (Mathison y Ross, 2008)

En 1989, en la llamada Cumbre Educativa de Charlottesville, los gobernadores de los 50 estados norteamericanos adoptaron un conjunto de metas en la perspectiva del año 2000. La tercera meta establecía que para esa fecha “los estudiantes americanos deberían terminar los grados 4°, 8° y 12° demostrando competencia en temas exigentes de inglés, matemáticas, ciencias, historia y geografía” (Mathison, 2008, pp. 8-9). En 1990 se establecieron procedimientos apoyados con fondos federales para avanzar hacia esas metas, y se crearon el National Education Goals Panel y el National Council on Education Standards and Testing.

La preocupación por la calidad educativa no fue exclusiva de Estados Unidos. Los hechos que agudizaron esa preocupación, en especial el lanzamiento del Sputnik, produjeron también reacciones que llevaron al surgimiento de las evaluaciones internacionales a lo largo de la segunda mitad del siglo XX. Aun si cada país contara con un sistema nacional de evaluación, la comparación de los resultados no seguiría, dadas las diferencias de los sistemas educativos en estructura,

currículos y calendarios escolares; además de las diferencias de contenido, grado de dificultad y enfoque de los instrumentos de evaluación mismos. Por ello, los trabajos pioneros de la International Association for the Evaluation of Educational Achievement (IEA) fueron notables (Postlethwaite, 1985; De Landsheere, 1994).

II. Proliferación de las pruebas a gran escala en el siglo XXI

La descentralización que ha caracterizado al sistema educativo norteamericano trajo consigo la consecuencia de que las evaluaciones a gran escala que se aplican en cada estado no se pueden comparar. El sistema National Assessment of Educational Progress (NAEP), establecido a fines de la década de 1960 (Walberg, 1990), ofrecía resultados confiables sobre el sistema en ciertas materias y grados en la escala nacional, pero no individual, de escuela o distrito, y ni siquiera estatal. Para dar resultados confiables a nivel de plantel se buscaron otras soluciones: la Prueba Nacional Voluntaria (Voluntary National Test), propuesta por Clinton, o un sistema de pruebas adaptativas computarizadas en internet, que propuso la Rand Corporation (Klein y Hamilton, 1999).

Al comenzar el año 2002, el presidente Bush promovió una nueva legislación educativa en el nivel federal, que se conoce con la expresión *No child left behind* (NCLB). Esta ley implicó cambios importantes en las políticas educativas en general, y en particular en lo que se refiere a la evaluación del rendimiento de los alumnos.

La Ley pretende modificar en un plazo bastante corto (12 años, es decir, para el 2014) la situación de la educación norteamericana, incluyendo las desigualdades que la caracterizan, con medidas, entre las que destaca reforzar los mecanismos de evaluación de la calidad educativa: todos los estados americanos deberán tener estándares de desempeño claros y sistemas estatales de evaluación alineados con ellos, así como aplicar anualmente pruebas de inglés, matemáticas y ciencias a todos los alumnos de los grados 4° a 8°.

La participación de los estados en las pruebas del NAEP será condición obligatoria para que puedan acceder a fondos federales para apoyar los programas de mejora educativa que la nueva legislación contempla. Los resultados de los alumnos en las pruebas estatales son el criterio para definir si la escuela logra los avances estipulados para recibir apoyos (*adequate yearly progress*, AYP), y puede ser cerrada si no lo consigue. Esto hace que las pruebas de alto impacto tengan las consecuencias que se discutirán más adelante.

Además de Estados Unidos, al comenzar el siglo XXI muchos países han puesto en marcha sistemas de evaluación similares, incluyendo a muchos de Europa, pero también del Asia Oriental y el cercano oriente, en especial en Israel; en países árabes comienzan a implantarse con apoyo de la UNESCO. En África destaca el South African Consortium for the Monitoring of Educational Quality. En América Latina, México y Costa Rica comenzaron a emprender evaluaciones a gran escala en educación básica desde los años de 1970 y 1980, pero sólo Chile

desarrolló un verdadero sistema de evaluación antes de 1990. En la última década del siglo XX y en la primera del XXI, casi todos los países lo han hecho (Martínez Rizo, 2009).

A nivel regional destaca el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE) de la Oficina Regional de la UNESCO para América Latina y el Caribe. En 1997 llevó a cabo un primer estudio en 3° y 4° grado de primaria, con participación de Argentina, Bolivia, Brasil, Chile, Colombia, Cuba, Dominicana, Honduras, México, Paraguay y Venezuela. En 2006 el LLECE llevó a cabo un segundo estudio en Argentina, Brasil, Chile, Colombia, Costa Rica, Cuba, Ecuador, El Salvador, Guatemala, México, Nicaragua, Panamá, Paraguay, Perú, República Dominicana y Uruguay (Martínez Rizo, 2008).

En el plano internacional, además de la ampliación de las evaluaciones de la IEA, las pruebas de la Organización para la Cooperación y el Desarrollo Económico (OCDE) se han extendido aún más, en lo que se conoce con las siglas PISA (Programme for Institutional Student Assessment). Estas pruebas, de enfoque no curricular y dirigidas a jóvenes de 15 años de edad, se aplicaron por primera vez en 32 países en el año 2000, y luego cada tres años. En 2009 participaron más de 60 países (Martínez Rizo, 2008).

En México, las pruebas estandarizadas comenzaron a usarse en la segunda mitad del siglo XX. En educación superior con pruebas de selección y en los niveles básicos del sistema educativo, con el uso de pruebas rudimentarias elaboradas por los maestros mismos o, más frecuentemente, por los supervisores, que las proporcionaban a las escuelas a su cargo. Desde la década de 1970, la Secretaría de Educación Pública comenzó a hacer evaluaciones a gran escala. Las primeras pruebas se aplicaron en 1972, para decidir la admisión de alumnos en secundaria. A fines de esa década se hicieron las primeras aplicaciones de pruebas a muestras de alumnos de primaria, con el proyecto llamado *Evaluación del rendimiento académico de los alumnos de 4º y 5º grados de educación primaria*. (Martínez Rizo, 2008).

La situación no avanzó mucho, sino hasta principios de la década de 1990, cuando la evaluación a gran escala recibió un importante impulso, por la conjunción de varias circunstancias. La principal ocurrió en 1992, con el Acuerdo Nacional para la Modernización de la Educación Básica (ANMEB), del que se derivaron la descentralización del sistema educativo y el programa Carrera Magisterial. Para asignar los estímulos de este programa, se decidió tomar en cuenta, entre otros elementos, los resultados de los alumnos, lo que trajo consigo la necesidad de aplicar cada año pruebas a gran cantidad de alumnos. La primera aplicación involucró a más de cuatro millones. Estas evaluaciones siguieron aplicándose hasta 2005, cuando el número de alumnos evaluados mediante ellas llegó a cerca de ocho millones. Un segundo elemento consistió en los programas compensatorios que el gobierno mexicano implementó a partir de 1991, con apoyo del Banco Mundial. Estos programas incluyeron un componente de evaluación, con la aplicación de pruebas a los alumnos beneficiados. A partir de 1994 se

buscó extender la aplicación de pruebas similares en todas las entidades, en forma permanente, mediante el proyecto denominado Estudio de Evaluación de la Educación Primaria (EVEP).

También en 1994, el ingreso de México a la OCDE mostró el interés de las autoridades por integrarse a la vida económica y política internacional, incluyendo la participación en evaluaciones educativas, como el TIMSS de la IEA, el LLECE y el proyecto PISA de la OCDE. En 1996 se emprendió un trabajo de definición de estándares curriculares y se desarrollaron evaluaciones en relación con ellos: las Pruebas de Estándares Nacionales, que se aplicaron por primera vez en 1998.

En el siglo XXI la evaluación educativa en México avanzó con la creación del Instituto Nacional para la Evaluación de la Educación (INEE) y con el desarrollo de nuevas iniciativas de la Secretaría de Educación Pública (SEP), en particular las pruebas censales denominadas Exámenes Nacionales del Logro Académico en Centros Escolares (ENLACE) (Martínez Rizo, 2008).

III. Consecuencias negativas de las pruebas y críticas al respecto

Sin considerar las de ingreso a las universidades, las pruebas usadas en niveles preuniversitarios en la mayoría de los estados norteamericanos a lo largo del siglo XX eran de bajo impacto: sus resultados no influían en las decisiones importantes que se tomaban con respecto a cada alumno, ni tampoco en las que tenían que ver con maestros y escuelas individuales. Esta situación comenzó a cambiar en la década de 1980, y la tendencia se acentuó en la de 1990, para culminar en las disposiciones de la Ley *No child left behind*, de 2002, con la que las pruebas a gran escala adquirieron un peso fundamental y sin precedentes en decisiones relativas a alumnos, maestros y escuelas.

En otros países ocurrió algo similar. El que las pruebas se aplicaran masivamente y sus resultados se difundieran mediante ordenamientos simples de escuelas, basados en los puntajes obtenidos en promedio por los alumnos, sin tener en cuenta el contexto en que opera cada una (*rankings* o *league tables*), volvía de alto impacto los resultados. Esto es verdad aun en ausencia de disposiciones legales precisas, que impliquen consecuencias oficiales fuertes basadas en esos resultados, como ha ocurrido en Estados Unidos, el Reino Unido o Chile.

Muchas de las críticas a las pruebas a gran escala provienen de personas que las rechazan en bloque, sin matices que tengan en cuenta sus variantes y los usos de sus resultados. En cambio, las críticas que se incluyen en los párrafos siguientes, vienen de personas conocedoras de los aspectos metodológicos relevantes de las pruebas estandarizadas y que, en general, son partidarias de un uso adecuado de ellas. A diferencia de los críticos radicales, lo que estos juicios cuestionan son usos de la evaluación que creen ilegítimos, porque no tienen en cuenta los alcances y las limitaciones de las pruebas. Por ello, tienden a hacer un uso abusivo de sus resultados, con consecuencias negativas que pueden ser serias (Martínez Rizo, en prensa).

A propósito del creciente peso que la evaluación basada en pruebas adquirió en Estados Unidos, antes de la Ley *No child left behind*, y de los riesgos que trajo consigo, un experto reconocido dijo que la tendencia se debía a la preocupación –en muchos casos fundada— de muchas personas respecto a la calidad de las escuelas, y que en ese contexto las pruebas adquirieron un peso predominante. Luego se refirió a las consecuencias negativas que trajo consigo esa importancia excesiva y mal enfocada que se estaba dando a las pruebas de rendimiento:

> 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 [traducción libre del autor] (Popham, 2001, Secc. Do you think the politicians know this?, párr. 2).

Popham (2001) dejaba claro que su postura no se refería a cualquier forma de usar pruebas para evaluar resultados educativos, sino a ciertas formas inapropiadas de hacerlo. Afirmaba, expresamente, que pruebas bien diseñadas y utilizadas adecuadamente pueden ser de gran valor para la educación:

> Está surgiendo en nuestro país una resistencia a cualquier tipo de pruebas. Pienso que esto no es sano. Creo que hay que usar pruebas bien construidas, que ayuden a los maestros a mejorar su enseñanza. Pienso también que el público tiene derecho a saber qué tan bien funcionan las escuelas. Por ello pienso que oponerse a cualquier tipo de pruebas es negativo para los alumnos. Tenemos que hacer buenas pruebas, que pueden ser una fuerza poderosa para mejorar la enseñanza, haciendo que los alumnos aprendan lo que deben aprender [traducción libre del autor] (Popham, 2001, Secc. I met this teacher…, párr. 1).

En un texto reciente, con la evidencia de que sus temores respecto a la extensión de las pruebas a gran escala sin las debidas consideraciones, se habían vuelto realidad, este experto precisa dos razones por las que una buena idea –conseguir que los alumnos alcancen altos niveles de competencia, con una educación basada en estándares— está teniendo las consecuencias que él anticipaba: por una parte, el exceso de contenidos que trae consigo una definición inadecuada de los estándares; por otra, el uso de pruebas inapropiadas, en concreto por su falta de sensibiidad a la instrucción (instructionally insensitive tests), como instrumentos para verificar el cumplimiento de los estándares (Popham, 2008).

La experiencia de la aplicación de la Ley *No child left behind*, después de su entrada en vigor, ha puesto en evidencia deficiencias importantes y consecuencias contraproducentes, sobre todo, para escuelas públicas. Varias proyecciones señalan que muy pocas de ellas podrán satisfacer las exigencias del avance anual

adecuado (*adequate yearly progress*), que establece la Ley; mientras que la gran mayoría (tal vez más de 95% en todo el país) deberán ser clasificadas como deficientes (*failing*) y enfrentar las consecuencias de ello, que pueden llegar teóricamente hasta su desaparición (Oakes y Lipton, 2007).

Un investigador destacado en el escenario psicométrico contemporáneo, Robert Linn, escribió también antes de la Ley *No child left behind*:

> Me veo llevado a concluir que, en la mayoría de los casos, los instrumentos y la tecnología no han estado a la altura de lo que esperaba de ellos la rendición de cuentas de alto impacto. Los sistemas de evaluación basados en pruebas, que son útiles para propósitos de monitoreo, pierden mucha de su confiabilidad y credibilidad para ello, cuando se les asocian consecuencias fuertes. Los efectos negativos inesperados de usos de alto impacto de la rendición de cuentas frecuentemente son más importantes que los efectos positivos que se buscaban [traducción libre del autor] (Linn, 2000, p. 16).

Contemporánea a la Ley es la cita siguiente, en la que tres estudiosos de la Rand Corporation precisan lo que probablemente explica las amplias expectativas que han llevado a usos inadecuados de las pruebas a gran escala y de alto impacto:

> 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 [traducción libre del autor] (Hamilton, Stecher y Klein, 2002, p. iii).

Muchas personas no tienen conciencia de la dificultad que implica obtener buenos resultados educativos con grupos de alumnos que provienen de un medio social desfavorable. En México es frecuente que dirigentes del sector empresarial vean con simpatía las estrategias simplistas a las que alude la cita anterior, pensando que las fallas de la escuela pública se podrían corregir fácilmente con escuelas privadas como las que atienden a sus hijos; pero ignoran que menos del 10% de los niños mexicanos, de condiciones privilegiadas, asisten a ellas. Probablemente por eso son frecuentes las opiniones de que bastará con aplicar pruebas masivamente, y tomar medidas correctivas simples, para que la calidad de la educación mejore sustancialmente.

En América Latina, hasta mediados de la década de 1990 los resultados de las pruebas que se aplicaban en educación básica no llevaban a decisiones que afectaran a individuos, como decidir si aprobar o reprobar a un alumno, asignar estímulos o tomar medidas correctivas que afecten a maestros o escuelas. Su impacto era bajo e incluso nulo, por la ausencia de difusión de los resultados. La excepción fue el Sistema de Medición de la Calidad de la Educación (SIMCE), de Chile, que desde sus inicios se definió como de alto impacto: su diseño censal se hizo con el propósito de contribuir a la introducción de cambios mayores en el sistema educativo, como su municipalización y relativa privatización. Los

resultados se han utilizado para decidir cuáles escuelas pueden recibir fondos públicos, en la forma de bonos individuales para sus alumnos.

Los desarrollos más recientes en nuestro *subcontinente* apuntan en una dirección similar a la observada en Estados Unidos: se tiende a pensar que aplicar pruebas censales, cuyos resultados permitan comparaciones directas y simples entre escuelas, facilitará tomar decisiones que llevarán a mejoras sustanciales a corto plazo. Además de Chile, Uruguay y México, otros países que están incursionando en la aplicación de pruebas censales son: Brasil, Colombia, Costa Rica, República Dominicana, Ecuador, El Salvador, Guatemala y Perú. El riesgo de que aparezcan consecuencias contraproducentes no es ya sólo teórico, sino real.

En México el balance de la evaluación tiene luces y sombras. En el lado positivo se deben mencionar los avances técnicos y la formación de especialistas de buen nivel; la creciente conciencia ciudadana del derecho a conocer los resultados de las evaluaciones, que contrasta con el hermetismo anterior; y el que algunas autoridades educativas, en el nivel federal y en algunos estados, comiencen a hacer uso de los resultados de las evaluaciones para toma de decisiones. En el lado negativo hay que contar el número ya excesivo y creciente de pruebas que se desarrollan y aplican, que pesa cada vez más sobre alumnos, maestros y escuelas; el predominio de la evaluación a gran escala que deben usar los maestros en el aula; el uso inapropiado, cada vez más frecuente, de los resultados y su excesivo peso en el diseño de las políticas públicas.

En síntesis, la proliferación de pruebas a gran escala va acompañada por el interés de que sus resultados se utilicen para sustentar decisiones de las que se deriven mejoras importantes para la calidad. Esta tendencia se relaciona con la de rendición de cuentas, y cobra sentido en el contexto de corrientes más amplias: búsqueda de transparencia en el manejo de los asuntos públicos; con frecuencia, desconfianza respecto de la educación pública y, en general, respecto a la gestión pública de los servicios. A ello debe añadirse la escasa cultura en la sociedad, en cuanto a evaluación educativa, no sólo entre el público general, sino también entre maestros y autoridades educativas, e incluso entre investigadores y especialistas. Esto lleva a esperar resultados casi milagrosos en las escuelas, gracias a la aplicación de pruebas, sin tener en cuenta sus alcances y límites.

Para que se concreten las perspectivas favorables que se asocian con las pruebas es necesaria una visión más completa de sus posibilidades. Hay que tener claro que las pruebas a gran escala tienen rasgos que limitan su capacidad para informar sobre muchos aspectos importantes de los que debe incluir el currículo. Lo anterior se ve considerablemente agravado en el caso de aplicaciones censales, sobre todo, si se pretende cubrir muchos grados y con mucha frecuencia. Además, dichas pruebas nunca podrán sustituir el trabajo del maestro.

El último punto tiene especial relevancia: sólo un buen maestro puede llevar a cabo la evaluación más importante de cada alumno. Una evaluación que incluya todos los aspectos del currículo y los niveles cognitivos más complejos, que tenga

en cuenta las circunstancias de cada niño, y se haga con la frecuencia necesaria para ofrecer retroalimentación oportuna para que el alumno pueda mejorar. Este tipo de evaluaciones son las que deben hacerse en cada aula regularmente, con acercamientos más finos que los que pueden emplearse a gran escala. Muchos maestros no tienen la preparación necesaria para hacer bien dicha evaluación, pero ninguna prueba a gran escala podrá ocupar su lugar. Por ello, habrá que ofrecer a los docentes los apoyos necesarios para que cumplan adecuadamente con su función evaluativa, viendo a las pruebas como uno de esos apoyos.

Aun si se hacen bien, las evaluaciones a cargo de los maestros tienen también limitaciones. En particular, sus resultados no son agregables, en el sentido de que permitan la construcción de medidas sintéticas ni pueden ofrecer información sobre la situación de conjuntos de grandes dimensiones, como son los sistemas educativos. Las pruebas a gran escala pueden ofrecer insumos valiosos para la toma de decisiones en diversos niveles del sistema, pero siempre que se les vea como complementos del trabajo de los maestros, y no pretendan sustituirlo.

Es fundamental, además, adoptar una perspectiva que no vea la evaluación como amenaza, sino como oportunidad de aprendizaje y de mejora. Los resultados de las pruebas, en vez de servir para hacer ordenamientos simples, podrían ayudar a detectar oportunamente alumnos en riesgo y escuelas que necesiten apoyo especial, y así para brindarlo oportunamente a unos y otras, en lugar de propiciar competencias estériles de las que se derivan consecuencias perversas.

IV. Evaluación formativa y evaluación en el aula

Si se tiene en cuenta que las evaluaciones que hacían los maestros a fines del siglo XIX y principios del XX consistían, en muchos casos, en recitaciones de tipo mecánico o memorístico, con las que un alumno mostraba lo que sabía (Oakes y Lipton, 2007), no sorprende que las pruebas estandarizadas fueran vistas como un avance, y se convirtieran en el referente que los maestros trataban de imitar, y para lo que se les preparaba.

Sin embargo, y además de sus ventajas, las pruebas a gran escala de enfoque normativo y preguntas de opción múltiple tienen también claras limitaciones, en especial en relación con la medición de niveles cognitivos complejos y por lo difícil que resulta controlar la influencia del contexto social de los alumnos en los resultados o, de otro modo, por su falta de sensibilidad a la instrucción. Por ello, desde las primeras décadas del siglo pasado se expresaron críticas a esas evaluaciones, cuestionamientos que arreciaron en la medida en que las pruebas a gran escala adquirieron mayor peso, como ha ocurrido en las últimas décadas.

En este apartado se verá el desarrollo de la alternativa a la evaluación a gran escala que son las evaluaciones a cargo de los maestros. La postura que ve a las pruebas a gran escala como complemento del trabajo docente, pero no como sustituto del mismo, parte de la idea de que la influencia de un buen maestro es

insustituible, tanto para que los alumnos aprendan, como para valorar el grado en que tal cosa ocurre, o sea, para evaluar.

Valorar el grado en que un alumno tiene los conocimientos y habilidades previstos al final de un ciclo escolar no es sencillo, si se quiere cubrir de manera suficiente las diversas materias o áreas del currículo y los temas de cada área o materia. La tarea se complica si se quiere conocer el avance del alumno –lo que es esencial para ofrecer retroalimentación–, ya que la evaluación deberá hacerse desde el inicio del ciclo escolar y en varios momentos del mismo, en forma permanente.

Esto último es básico si se quiere que la evaluación sea útil no sólo para detectar el resultado final de un proceso educativo (lo que se conoce como evaluación sumativa), sino, sobre todo, para contribuir a que el proceso de aprendizaje mejore en toda su extensión, a lo que alude la expresión evaluación formativa. Si se trata de valorar el avance cotidiano de dos o tres decenas de alumnos, y se quiere tener información sobre las circunstancias personales, familiares y sociales de cada uno, para tenerla en cuenta en el momento de tomar decisiones importantes para el futuro de cada uno de ellos, la tarea evaluativa se vuelve compleja.

Como se mencionó antes, desde los inicios del desarrollo de las pruebas a gran escala, algunos de sus promotores más lúcidos, como Tyler, señalaban que también ese tipo de evaluaciones debían verse como parte del proceso de enseñanza-aprendizaje, pero prevaleció un enfoque que en realidad las manejaba como un elemento adicional que sólo tenía lugar al final del mismo.

La distinción entre la evaluación final y la que tiene lugar a lo largo del proceso, entre *evaluación formativa* y *sumativa*, es reciente.

La Teoría Clásica de los Tests y las pruebas de rendimiento a gran escala de diseño tradicional se desarrollaron 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* (Shepard, 2006).

En la medida en que se identifican y exploran los procesos mentales se abren horizontes vastos y atractivos tanto para la pedagogía como para las metodologías de evaluación del aprendizaje, en especial para las que pretendan servir a propósitos *formativos*, aportando elementos para que maestros y alumnos modifiquen sus acciones en consecuencia, para alcanzar mejores resultados. En

este sentido, un importante trabajo sobre la evaluación en aula apunta elementos importantes en lo que se refiere al potencial formativo de las evaluaciones:

> 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 (Shepard, 2006, p. 626).

El interés por la evaluación en aula con propósitos formativos se relaciona con la creciente conciencia de las limitaciones de las pruebas convencionales para tales fines, y con avances paralelos debidos a los expertos en áreas de contenidos curriculares que, tanto por el rechazo de los efectos de las pruebas usadas para rendición de cuentas, como por los profundos cambios en las concepciones del aprendizaje y del manejo adecuado de los contenidos, comenzaron a desarrollar alternativas a las pruebas para su uso en el aula (Shepard, 2006).

Como se ha dicho, muchos maestros no tienen la competencia necesaria para manejar evaluaciones en aula que sean superiores a las de gran escala en lo relativo a su potencial para retroalimentar su trabajo y el de sus alumnos. Por ello, desde 1989, Silver y Kilpatrick (como se cita en Shepard, 2006, p. 627) 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 [traducción libre del autor].

Un trabajo muy reciente presenta un interesante resumen de la forma en que ha evolucionado la definición de evaluación formativa (Brookhart, 2009).

La idea original que distingue la información que se usa para mejorar algo que está en proceso, en oposición a la que sirve para valorar el resultado final, la propuso Michael Scriven, en 1967. Este autor se refería a la evaluación del currículo y de programas educativos. Pronto otros autores cayeron en la cuenta de la importancia de esa distinción que, aunque hoy parece obvia, no se había manejado explícitamente antes del trabajo seminal de Scriven.

En 1971 apareció el libro de Bloom, Hasting y Madaus, que popularizó las nociones de evaluación formativa y sumativa, aplicadas ya al aprendizaje de los alumnos. En esa obra se precisan las diferencias de las evaluaciones que se usan para apoyar decisiones instruccionales, distinguiendo los propósitos formativos y los sumativos, así como los de ubicación y diagnóstico. Brookhart (2009) subraya que este trabajo añade al concepto de Scriven un elemento importante: que, además de *ofrecer información sobre el proceso de aprendizaje y no sólo sobre sus resultados finales*, lo que la evaluación formativa aporta *puede servir a los maestros para que tomen mejores decisiones instruccionales*. Puede

añadirse que Bloom puso en práctica sus ideas con el sistema de enseñanza Mastery Learning, basado en el modelo de aprendizaje de Carroll.

La noción se desarrolló con Sadler (1989), para quien no sólo el docente puede usar resultados de evaluación formativa, sino también los alumnos. Con este autor el calificativo de *formativo* se aplica al sustantivo que designa la evaluación del aprendizaje de los alumnos (*assessment*) y no, como en Scriven y Bloom, al de *evaluation*, que se refería a currículos y programas (Brookhart, 2009).

Un paso más en la precisión de la idea de evaluación formativa se da cuando se destaca la importancia de los aspectos afectivos de la retroalimentación que se da a los alumnos. Hasta entonces el énfasis se ponía en los aspectos cognitivos. Más recientemente autores como Black y Wiliam (1998), Stiggins (2008) y Brookhart (2009), subrayan esta dimensión.

El último autor citado señala que, hasta hace poco, se consideraba aceptable que sólo unos alumnos alcanzaran los objetivos de aprendizaje, mientras muchos no lo lograban. El papel de la evaluación era distinguir unos de otros, y los criterios para valorar la calidad de las evaluaciones eran su validez y su confiabilidad. Hoy se espera de las escuelas que consigan que todos los alumnos alcancen los niveles de competencia necesarios, lo que obliga a reflexionar sobre las formas apropiadas para evaluar el aprendizaje en este nuevo contexto, lo que tiene que ver con el impacto emocional de la evaluación sobre los alumnos, Stiggins dice:

> Desde los primeros grados, algunos alumnos... obtienen altos puntajes en las evaluaciones y reciben altas calificaciones. El efecto emocional es que se ven a sí mismos como capaces de aprender, y se sienten cada vez más confiados... Otros, en cambio, obtienen puntajes bajos en las pruebas y reciben calificaciones malas. Esto los lleva a dudar de su capacidad. La falta de confianza en sí mismos los priva de las reservas emocionales para correr el riesgo de seguir intentando... Si un alumno se rinde y deja de esforzarse, o incluso si abandona la escuela, eso es visto como un problema del alumno, no de sus maestros o de la escuela. La responsabilidad de ésta es ofrecer oportunidades de aprendizaje, si los alumnos no las aprovechan, no es responsabilidad del sistema [traducción libre del autor] (2008, p. 7).

Stiggins (2008) añade que la importancia del cambio de paradigma que implica centrar la atención en los alumnos, como usuarios privilegiados de los resultados, teniendo en cuenta el impacto afectivo de las evaluaciones, no se puede exagerar:

> Durante décadas los expertos en la mejora escolar han cometido el error de pensar que los adultos del sistema son los usuarios más importantes de las evaluaciones. Hemos creído que si los adultos toman mejores decisiones en lo relativo a la enseñanza, las escuelas se volverán más eficaces… Pero esta visión pierde de vista la realidad de que los alumnos pueden ser tomadores de decisiones de aprendizaje más importantes que los adultos... Si un alumno decide que cierto aprendizaje está fuera de su alcance o que el riesgo de fracaso público es demasiado grande o amenazador, entonces, hagamos lo que hagamos los adultos, el aprendizaje termina. Por ello la pregunta fundamental para maestros y

> directores de escuela es: ¿qué podemos hacer para ayudar a que los alumnos respondan en forma productiva las preguntas anteriores, que los mantengan con esperanza de que el éxito está a su alcance si persisten en el intento? [traducción libre del autor] (p. 8)

Cada una de las etapas del desarrollo de la noción de *evaluación formativa* ha aportado algo sustantivo: la idea original de Scriven, que distingue la evaluación al final o durante el proceso; la aplicación explícita de la noción a la evaluación del aprendizaje, y no sólo del currículo o programas, por Bloom; la identificación de los alumnos como destinatarios clave de la información, con Sadler; y, finalmente la atención a la dimensión afectiva, con Brookhart, Black y Wiliam y Stiggins.

V. Conclusión

Hacer evaluación formativa, en el aula o en un nivel más amplio, no es sencillo, pero si no se consigue dar ese giro a la evaluación, su utilidad como herramienta de mejora será reducida. Por ello, dar a los maestros elementos que les permitan orientar su trabajo de evaluación en sentido formativo es importante y complejo.

El giro que está tomando la evaluación en nuestro sistema educativo no parece ir en la dirección correcta. Sin desconocer el lado positivo de los avances que se han dado en los últimos años en México, en evaluación educativa, parece que ha llegado el momento de advertir lo anterior. Esta llamada de atención podrá ser más productiva si, al mismo tiempo, se propone una alternativa mejor. La dirección alternativa no es otra que la de un sistema de evaluación que combine de manera más equilibrada evaluaciones a gran escala parsimoniosas y consistentes, con un rico trabajo de evaluación formativa en aula a cargo de los maestros.

En estos últimos párrafos se desarrolla esta idea, siguiendo de nuevo a Stiggins (2008) en su texto, titulado significativamente *Un manifiesto por la evaluación: Llamada por el desarrollo de sistemas de evaluación equilibrados*.

> Un manifiesto es la expresión pública de intenciones, creencias, opiniones o propuestas de políticas a favor de cierta acción política o social. Con frecuencia tan ardientes expresiones se oponen a los valores y prácticas convencionales o dominantes. He decidido difundir este manifiesto porque estoy convencido de que hemos llegado a un punto decisivo en la evolución de nuestros sistemas educativos, en el que debemos reevaluar, redefinir y rediseñar el papel de la evaluación en el desarrollo de escuelas eficaces. La tarea a emprender es tan importante que exige una urgente acción pedagógica, social y política [traducción libre del autor] (p. 2).

En las últimas páginas del texto de Stiggins (2008), este autor explica en qué consiste su manifiesto, en términos de la evaluación total como solución:

> Hoy entendemos mucho mejor que antes cómo usar productivamente la evaluación. Debemos sustituir los pasados sistemas, marcadamente

> desequilibrados, por otros que satisfagan las necesidades de información de todos los usuarios: sistemas que, a la vez, verifiquen el aprendizaje y lo apoyen, desde el aula hasta la sala de juntas de las autoridades. Para conseguir el equilibrio tan necesario y tan largamente ausente, debemos implementar prácticas de evaluación en aula que se apoyen en una gama de aproximaciones a la calidad usadas estratégicamente de manera que mantengan la fe de los alumnos en sí mismos… es tiempo de sustituir la intimidación de la rendición de cuentas como principal motivación, por la promesa del éxito académico para todos los aprendices, como esa fuerza motivacional. El miedo a veces funciona, pero sólo motiva a los que tienen esperanza de tener éxito. Desafortunadamente, cuando se trata de aprender la desesperanza siempre pesa más que la intimidación. Una evaluación en aula efectiva puede y debe servir para promover la esperanza en todos los alumnos [traducción libre del autor] (p. 10).

Stiggins (2008) señala que hoy estamos en condiciones de modificar los sistemas de evaluación en la dirección correcta; que, gracias a las investigaciones que se han llevado a cabo durante las últimas dos décadas, se cuenta con lo necesario para poner en marcha actividades formativas para maestros que les den la competencia necesaria para hacer buenas evaluaciones, y comenta que las escuelas de educación deberían caminar en esta dirección.

La situación del sistema educativo mexicano es similar, pero más grave. Por ello la necesidad de equilibrar nuestro sistema de evaluación es aún más apremiante. Stiggins termina su alegato diciendo que tenemos lo necesario, y que la única pregunta que necesita respuesta es una que también debe plantearse en México:

> ¿Tendrán educadores y diseñadores de políticas la oportunidad de aprender a evaluar productivamente? Históricamente la respuesta inequívoca ha sido: no. A consecuencia de ello, el inmenso potencial de la evaluación para apoyar el aprendizaje se ha desaprovechado, sin duda incluso ha pasado desapercibido en los niveles más altos de la toma de decisiones. No tiene por qué ser así. Está a nuestro alcance adoptar una nueva visión de una evaluación de excelencia, que libere la fuente de confianza, motivación y potencial de aprendizaje que hay en cada alumno. Es tiempo de hacerlo [traducción libre del autor] (Stiggins, 2008, p. 12).

Al revisar la literatura sobre el tema de la evaluación en aula con propósitos formativos, se hace evidente el creciente interés que despierta en los medios educativos. Las referencias eran raras a fines de la década de 1980; aumentaron a lo largo de la de 1990, en especial en su segunda parte; y se volvieron numerosas en lo que va del siglo XXI. Un trabajo aún no publicado con dos centenares de referencias sobre el tema, en su mayoría de la última década, está a la disposición de las personas interesadas (Martínez Rizo, 2009).

Referencias

Baker E. L. y Choppin, B. H. (1990). Minimum competency testing. En H.J Walberg y H. J. Haertel (Eds.). (1990). *The International Encyclopedia of Educational* Evaluation (pp. 499-502). Oxford-Nueva York: Pergamon Press.

Black, P. y Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7-74.

Bloom, B. S, Hastings, J. T., Madaus, G. F. (Eds.). (1971). *Handbook on formative and summative evaluation of student learning*. Nueva York: McGraw-Hill.

Brookhart, S. M. (2009). Editorial. *Educational Measurement: Issues and Practice, 28* (1), 1-2.

De Landsheere, G. (1996). *La investigación educativa en el mundo* (Trad. G. A. Gallardo Jordán). México: Fondo de Cultura Económica. (Trabajo original publicado en 1986).

Hamilton, L. S., Stecher, B. M. y Klein S. P. (Eds.). (2002). *Making sense of test-based accountability in education*. Santa Monica, CA: Rand Corporation.

Klein, S. P. y Hamilton, L. (1999). *Large-scale testing. Current practices and new directions.* Santa Monica, CA: Rand Education.

Linn, R. (2000). Assessments and accountability. *Educational Researcher, 29* (2), 4-16.

Martínez, Rizo, F. (2008). *Las evaluaciones educativas en América Latina* (Serie: Cuadernos de Investigación, No. 32). México: Instituto Nacional para la Evaluación de la Educación.

Martínez Rizo, F. (2009). *Marco de referencia para el proyecto "El uso formativo de la evaluación para mejorar el aprendizaje"*. Manuscrito no publicado, Universidad Autónoma de Aguascalientes-Instituto de Investigación, Innovación y Estudios de Posgrado de la Educación.

Martínez Rizo, F. (en prensa). Assessment in the context of educational policy: The case of Latin American Countries. En E. Baker, B. McGaw y P. Paterson (Eds.), *International Encyclopedia of Education* (3a ed.). Oxford-Nueva York: Elsevier.

Mathison, S. (2008). A short history of educational assessment and standards-based educational reform. En S. Mathison y E. W. Ross (Eds.), *The nature and limits of standards-based reform and assessment* (pp. 3-14.). Nueva York: Teachers College Press.

Oakes, J. y Lipton, M. (2007). *Teaching to change the world* (3ª ed.). Nueva York: McGraw Hill.

Phelps, R. P. (2000). Trends in large-scale testing outside the United States. *Educational Measurement: Issues and Practice, 19* (1), 11-21.

Popham, W. J. (2001). Interview: James Popham. *Frontline.* Consultado el 7 de octubre de 2009, en: http://www.pbs.org/wgbh/pages/frontline/shows/schools/interviews/popham.html

Popham, W. J. (2008). Standards-based education: Two wrongs don't make a right. En S. Mathison y E. W. Ross (Eds.), *The nature and limits of standards-based reform and assessment* (pp. 15-25). Nueva York: Teachers College Press.

Postlethwaite, N. (1985). International association for the evaluation of educational achievement (IEA). En T. Husén y N. Postlethawite (Eds.), *International encyclopedia of education* (pp. 2645-2646). Nueva York: Pergamon Press.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

Scriven, M. (1967). The methodology of evaluation. En R. Tyler, R. Gagne y M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago, IL: Rand McNally.

Shepard, L. A. (2006). Classroom assessment. En R. L. Brennan (Ed.), *Educational measurement* (4a. ed., pp. 623-646). Westport, CT, Estados Unidos: Praeger.

Stiggins, R. (2008). *Assessment manifesto: A call for the development of balanced assessment systems.* Portland, Estados Unidos: ETS Assessment Training Institute.