Evaluación educativa y pruebas estandarizadas. Elementos para enriquecer el debate

Felipe Martínez Rizo.

Resumen

Al presentar algunos aspectos de la historia norteamericana y mundial de las evaluaciones en gran escala, este artículo muestra que los problemas que enfrentan las instituciones que se dedican a esa tarea en nuestro país no son inéditos, sino que se han presentado en otros lugares, que hay soluciones probadas para algunos, y que en el último medio siglo los avances en este campo han sido muy importantes. En México las pruebas de opción múltiple de buena calidad técnica comienzan apenas a extenderse y se considera novedosa la utilización de la teoría clásica de las pruebas en su diseño y la interpretación de sus resultados, sin embargo, tiene lugar enfrentando resistencias y rechazos en un contexto en el que tanto los partidarios de las pruebas como sus críticos parecen desconocer los avances que permiten corregir satisfactoriamente muchas de las limitaciones reales de las pruebas tradicionales.

Palabras clave: evaluación del aprendizaje, psicopedagogía, pruebas psicométricas.

Abstract

This article includes some aspects of the US and world history in terms of great scale evaluation experiences. Problems faced by Mexican institutions devoted to this task are not unique, they have been faced in other places, and, moreover, according to some people proven solutions exist since advancements made in this area during the past 50 years have been significant. Good technical quality multiple choice exams in Mexico are starting to expand, and the application of the classical theory in their design and result interpretation is seen as quite a novelty still. Nevertheless, reluctance exists since both, critics and supporters, seem to ignore the great advancements which allow for the satisfactory correction of many real limitations of traditional exams.

Key words: learning evaluation, psychopedagogy, psychometric tests.

Introducción

La evaluación del aprendizaje de los alumnos es un elemento tan antiguo y omnipresente en las instituciones escolares que seguramente puede considerarse esencial a todo sistema educativo. La evaluación, sin embargo, puede efectuarse de maneras muy diversas. En México esta función ha correspondido a los maestros, generalmente en forma individual. La aplicación de procedimientos sistemáticos, iguales o equivalentes, a grandes números de estudiantes, no ha sido usual en nuestro medio; si bien desde los años sesenta comenzaron a aplicarse pruebas en gran escala para el ingreso a la UNAM, esto no dejó de ser una excepción y, hasta muy recientemente, no se utilizaban técnicas que aseguraran la equivalencia de los instrumentos aplicados de un año a otro.

La SEP comenzó a aplicar pruebas estructuradas de aprendizaje a muestras nacionales de niños de primaria desde los años setenta, pero también sin cuidados técnicos para asegurar la equivalencia, y ha sido hasta los noventa, con la ampliación de estas acciones para la educación básica, y la creación del CENEVAL para la media superior y la superior, cuando este tipo de evaluaciones ha comenzado a extenderse en nuestro país. Como se sabe, esto ha sido seguido por fuertes manifestaciones de rechazo por parte de diversos sectores tanto estudiantiles como magisteriales.

Seguramente ese rechazo está motivado por una combinación de diversas razones de tipo académico —discrepancia sobre los fundamentos teóricos y metodológicos de las evaluaciones— y de naturaleza política, como temores, fundados o no, de que cierto tipo de pruebas resulten contraproducentes desde un punto de vista pedagógico, o inequitativas para determinado tipo de alumnos como las mujeres, los indígenas o los estudiantes de medio rural o urbano marginal.

La desconfianza se debe también al desconocimiento de la naturaleza de este tipo de trabajos. Fallas no atribuibles a las pruebas ni a los organismos que las elaboran, como usos inadecuados de los resultados, contribuyen al malestar. Un ejemplo de esto último es la utilización de los puntajes obtenidos por los aspirantes a ingresar a una institución de educación superior (IES) como único elemento de juicio para tomar la decisión respectiva, contra todas las recomendaciones técnicas, que señalan que dichos puntajes deberán combinarse con el promedio de bachillerato y, eventualmente, con otros elementos, para sustentar más adecuadamente tan delicadas decisiones. Conviene recordar, a este respecto, la advertencia que se incluía en el primer manual de la prueba más conocida internacionalmente, el *Scholastic Aptitude Test* , hace ya 75 años:

> El estado actual de los esfuerzos de los hombres por medir o, de alguna manera, apreciar el valor o mérito de otros hombres, evaluar los resultados de su educación, o estimar sus posibilidades o potencial, no garantiza certeza alguna de las predicciones que se hagan al respecto... Este nuevo test que ahora se ofrece a través del *College Entrance Examination Board* puede ayudar a resolver unos cuantos problemas que suscitan perplejidad, pero debe considerarse solamente un elemento adicional para ello. Poner demasiado énfasis en los puntajes obtenidos en las pruebas es tan peligroso como dejar de evaluar cualquier puntaje o rango, junto con otras medidas y estimaciones a las que complementa. (Brigham *et al.* , 1926)

Contra lo que sucede en realidad, se oye también que la utilización de estos instrumentos en México se da justamente cuando comienzan a abandonarse en los lugares en que su uso inició más tempranamente. Por todo lo anterior, una presentación histórica del desarrollo de las evaluaciones educativas en gran escala podrá contribuir a una mejor comprensión de su naturaleza y, en consecuencia, a la adopción de posturas más informadas al respecto. Esto es lo que busca el texto que se presenta a continuación que, dada la importancia del tema, especialmente para las IES, como muestra la polémica alrededor del CENEVAL, parece de especial actualidad.

Desarrollos teóricos y técnicos en el campo de las pruebas

Las bases de la teoría de la medición fueron puestas desde el siglo XVIII por Laplace y Gauss. La aplicación de estas ideas al campo educativo comenzó desde el XIX, distinguiéndose países como Alemania, Inglaterra, los Estados Unidos y, en menor medida, Francia y las regiones francófonas de Suiza y Bélgica. A partir de los primeros años del siglo XX se desarrolla la metodología que se conoce ahora como teoría clásica de las pruebas (*classical tests theory*), a partir de la teoría de la confiabilidad y el modelo estadístico de las puntuaciones, con las nociones de puntaje verdadero, error de medida y confiabilidad de la prueba (*true score, measurement error & test reliability, Cfr.* Keeves, 1997: 707). Los trabajos pioneros fueron los del inglés Charles Spearman, entre 1904 y 1913, y la obra del norteamericano Edward L. Thorndike *An introduction to theory of mental and social measurement* , publicada en Nueva York, también en 1904 (Martínez Arias, 1995: 40).

Según Du Bois, a partir de la cuarta década del siglo XX, con la aparición de publicaciones como *Psychometrika* (1935) y *Educational and Psychological Measurement* (1941) la teoría de los tests se separa de la evaluación y la psicología diferencial y, en su versión clásica, puede considerarse completa con la aparición del libro *Theory of mental tests* de Gulliksen, en 1950 (Martínez Arias, 1995: 42). La primera edición del *Mental Measurement Yearbook* es de 1938 y la *Psychometric Society* fue fundada a iniciativa de Thurstone en 1935. En 1946 Stevens formuló la clasificación ahora canónica de los niveles de medición nominal, ordinal, de intervalo y de razón (De Landsheere, 1996: 68).

Durante la segunda mitad del siglo XX el avance no se detuvo. En lo estrictamente psicométrico, los conceptos básicos mantuvieron su vigencia (confiabilidad, validez etc.), pero la manera relativamente ingenua en que los aplica la teoría clásica es progresivamente enriquecida por los planteamientos más sofisticados de dos importantes teorías: la de la generalización (*generalizability*) y la de respuesta al item (*item response theory*).

La teoría de la generalización, desarrollada por Cronbach y colaboradores, es una extensión de la teoría clásica, que atiende en forma más satisfactoria la problemática de la confiabilidad, substituyéndola por la noción de generalizabilidad: en lugar del concepto de puntaje verdadero se usa el de puntaje del universo (*universe score*), y en lugar de manejar el error de medición en forma global se identifican fuentes posibles de error o facetas y se detecta su influencia gracias a técnicas estadísticas como el ANOVA. El trabajo inicial, de 1963, precedió una década a la versión madura de la teoría (Cronbach,1972).

La teoría de respuesta al *item* fue esbozada por Lazarsfeld en sociología y por Lord en psicología. Tras la difusión de las computadoras se produjeron los modelos estadísticos de Birnbaum en los Estados Unidos (1957-58) y del danés Rasch, en 1960. La obra que difunde una teoría ya madura es la de Lord y Novick, de 1968: *Statistical theories of mental test scores* . La teoría de respuesta al item, o de las curvas características de un item, "intenta dar una fundamentación probabilística al problema de la medición de rasgos y constructos no observables (rasgos latentes), considerando al item, y no al puntaje global, como la unidad básica de análisis" (Martínez Arias, 1995: 237-243).

Las últimas décadas han visto otras novedades en el campo de las pruebas. Unas han sido sólo la precisión de nociones clásicas, como la de validez, que se concebía simplemente como "el grado en que un test mide lo que dice medir" (Garret, 1937), y luego se volvió cada vez más compleja, distinguiéndose validez de contenido, de criterio concurrente o predictivo y de constructo (Martínez Arias, 1995: 329-335) para desembocar en una concepción unitaria, con varias fuentes de evidencia de la validez. Pero además ha habido diversas innovaciones metodológicas y prácticas importantes:

- Técnicas para valorar el sesgo de un item o instrumento, en relación con variables como género o grupo étnico de los sustentantes: análisis de tablas de contingencia y de varianza, diagramas de dispersión de coeficientes delta, dificultades transformadas y procedimientos basados en la teoría de respuesta al item, *differential item functioning* (Martínez Arias, 1995: 577-612).

- Tests que miden el grado en que un sustentante alcanza un nivel de "maestría" previamente definido, en lugar de determinar su posición en relación con los demás sujetos: tests referidos a criterio (*criterion referenced tests*), según la terminología introducida por R. Glaser en 1963, desarrollados luego por Popham, Husek y Hambleton (Martínez Arias, 1995: 653-693).

- Desarrollo de *items* que no se limiten a presentar alternativas estructuradas de respuesta entre las que debe escoger el sustentante (opción múltiple, falso-verdadero), sino que requieran de la elaboración de la respuesta, acercándose a las tradicionales preguntas tipo ensayo, pero con modalidades que permitan una calificación objetiva, con procedimientos sistemáticos de jueceo.

- Pruebas de ejecución o desempeño, en las que las acciones que requiere el responder la prueba se aproximan lo más posible a las que deberá ejecutar una persona al desempeñar en la vida real una actividad determinada, a lo que se refiere la expresión *authentic testing* .

- Pruebas adaptativas por computadora, que no son simplemente pruebas tradicionales en un soporte moderno, sino que permiten ajustar el conjunto de items al nivel de conocimientos de cada sustentante, presentándole sucesivamente preguntas de un grado de dificultad que depende de sus respuestas anteriores, con lo que es posible incrementar la eficiencia de la prueba.

- Aprovechamiento del análisis factorial para identificar factores subyacentes a varios *items* y facilitar la definición de constructos, y el uso intencional de items multidimensionales.

- El "muestreo matricial" (*matrix sampling*) que permite aplicar pruebas que cubran mejor ciertos dominios, cuando no se desea obtener resultados confiables en el nivel individual sino grupal.

- Las *adaptaciones* , o variaciones controladas de los procedimientos de aplicación de una prueba, para atender las condiciones particulares de ciertos sustentantes, como los afectados por determinadas discapacidades (*accomodations*).

Estos y otros avances permiten hablar de una nueva generación de pruebas, muy distintas de las de los años cincuenta, que eran típicamente preguntas de opción múltiple aplicadas en forma controlada, cuyos resultados se analizaban con la teoría clásica. Ahora hay una rica gama de pruebas (adaptativas, de ejecución, libres de sesgo) que atienden necesidades de sustentantes especiales, que se analizan con base en las teorías de respuesta al item, la generalizabilidad y las aplicaciones psicométricas de los sofisticados avances de las estadística multivariada, como los modelos de ecuaciones estructurales, los de variables latentes y los diseñados para manejar variables categóricas.

Las pruebas del *College Board*

Un aspecto particular de la historia de la medición educativa en Estados Unidos, es el que se refiere al organismo más importante en lo relativo al desarrollo de pruebas para el nivel superior del sistema educativo, tanto para el ingreso a las universidades, como para el egreso de ellas.

El nacimiento del *College Board* , en noviembre del año 1900, tuvo lugar en un momento en que el incremento de jóvenes que terminaban la educación media superior, con la proliferación de *colleges* , que tenía lugar en ese momento en los Estados Unidos, hacía muy complejos los procesos de selección para entrar a una institución. Fuess señala que la fundación del *College Board* fue el primer intento organizado de "introducir la ley y el orden en una anarquía educativa que, a fines del siglo XIX, había llegado a ser exasperante, sin duda casi intolerable, para los directores de escuelas" (Citado en Donlon, 1984: 1). En esa época había "un consenso preocupantemente reducido entre los *colleges* en cuanto al tipo de preparación en ciertas áreas de contenido y en cuanto a los estándares de desempeñoque debían pedirse a los aspirantes (Donlon, 1984: 1).

Las primeras pruebas del *College Board* eran muy distintas de las actuales: pruebas de tipo ensayo en nueve áreas, cuya equivalencia se aseguraba aplicándolas en forma simultánea y asegurando la uniformidad de contenido, de condiciones de administración y de la calificación de las respuestas (Donlon, 1984: 1). Las pruebas fueron elaboradas por comités de maestros reconocidos; fueron aplicadas por primera vez en 1901, a 973 aspirantes, en 69 lugares; y fueron calificadas por comités de revisores en la Universidad de Columbia. En 1902 se aplicaron pruebas a 1,362 aspirantes a ingresar a alguna institución de educación superior y para 1910 el número llegó a 3,731. Muchos *colleges* y *high schools* veían al *College Board* como una amenaza para su autonomía; una consecuencia de esto fue el abandono de las etiquetas calificativas asociadas inicialmente a los puntajes numéricos, que eran Excelente para puntajes de 90 a 100; Bueno de 75 a 89; Dudoso de 60 a 74; Pobre de 40 a 59; y Muy Pobre menos de 40 (Donlon, 1984: 2).

Fue hasta 1925 cuando, gracias al desarrollo de las técnicas psicométricas, el *College Board* decidió desarrollar pruebas de aptitud, en contraposición a conocimientos, buscando ir más allá de la memorización de datos aislados, y acercándose a la evaluación de habilidades intelectuales básicas de tipo general. La prueba que todavía hoy se llama *Scholastic Aptitude Test* (SAT) se gestó a partir de abril de 1925, y vio la primera luz el 23 de junio de 1926, cuando se administró a 8,040 sustentantes. Las nueve subpruebas iniciales (definiciones, aritmética, clasificaciones, lenguaje artificial, antónimos, series, analogías, inferencias lógicas y lectura) se redujeron a siete en 1928 y a seis en 1929, agrupadas en dos secciones, de aptitud verbal y numérica (Donlon, 1984: 2).

Como no había técnicas de igualación de versiones, el constatar que el porcentaje de sustentantes que obtenía puntajes aprobatorios variaba bastante de año en año hizo pensar que lo que cambiaba en realidad era el grado de dificultad de la prueba, y no el nivel de los sustentantes que se suponía más estable. Por ello se decidió establecer una proporción fija de aprobados, ajustando las puntuaciones de los sustentantes (Donlon, 1984: 3).

En 1937 comenzó a hacerse una segunda aplicación anual que, a diferencia de la tradicional que seguía siendo de tipo ensayo, consistía en pruebas íntegramente compuestas por preguntas de opción múltiple. Ambas se calificaban en una escala con media de 500 puntos y desviación estándar de 100. La igualación de versiones y el cuidado de la estabilidad de la prueba comenzó a hacerse en 1941, cuando se estandarizaron los puntajes del SAT con la población de 10,654 sustentantes de la aplicación de abril de dicho año. A partir de 1942 todas

las aplicaciones usaron exclusivamente preguntas de opción múltiple. El número de sustentantes anuales llegó desde fines de los años sesenta al millón y medio de personas, cifra que se ha mantenido desde entonces (Donlon, 1984: 3-8).

Desde sus inicios, el *College Board* estableció un Comité Revisor para supervisar el desarrollo de sus pruebas. Con los avances de la psicometría fueron sistematizándose y ampliándose las medidas de control de calidad. En 1971 se publicó la primera edición del Manual Técnico del SAT, que no era un documento como el de 1926 y otros posteriores, en los que se daba información al usuario sobre la manera de utilizar correctamente la prueba para propósitos de orientación o selección. A partir del de 1971, el manual es una obra que sintetiza trabajos técnicos y resultados de investigaciones, ofreciendo al especialista "toda la información necesaria para una evaluación técnica exhaustiva de la prueba", con estudios dirigidos a analizar la validez de constructo, de contenido y predictiva, estadísticas descriptivas de los sustentantes, síntesis de trabajos sobre el efecto del entrenamiento y el posible sesgo de la misma y análisis similares (Donlon, 1984).

Otros datos muestran el avance en la calidad metodológica del SAT: el tamaño de la muestra de sustentantes para probar un item pasó de 370 en 1961 a 2,000 en 1975; la proporción de irregularidades en la aplicación es menor al 0.1%; la de errores en el proceso de lectura mecánica es prácticamente 0. Desde 1953 se hace corrección por adivinación. A partir de 1982 comenzó a usarse la Teoría de Respuesta al *item* . La nueva versión del SAT que entró en operación a mediados de los noventa incluye *items* de respuesta construida. Otros cambios se debieron a decisiones legales, en particular las aprobadas en 1980 en Nueva York, que obligan a publicar cada prueba después de su aplicación, a entregar a cada sustentante las respuestas correctas y sus propias respuestas, a entregar a la autoridad educativa estatal copia de todos los estudios relacionados con el SAT, así como dar información sobre lo que mide la prueba, sus limitaciones y la forma de calificarla.

La evaluación educativa en gran escala en el mundo en la actualidad

La difusión de la evaluación en el mundo

Entre los que se oponen a las pruebas en gran escala, no es raro en México oír opiniones en el sentido de que esas pruebas están siendo abandonadas en países avanzados, incluidos los Estados Unidos, supuestamente por haberse extendido la conciencia de sus insuperables limitaciones. Curiosamente, los opositores norteamericanos a las pruebas utilizan el mismo argumento, afirmando que están siendo abandonadas en otros países y atribuyendo a su fuerte presencia las fallas de las escuelas estadounidenses. Phelps (2000: 11) menciona cinco publicaciones americanas recientes en ese sentido y señala que las afirmaciones referidas no presentan sustento sólido, sino que se limitan a afirmar su posición señalando, por ejemplo, que Bélgica, Grecia, Portugal y España han eliminado ese tipo de exámenes. Si se sabe que ha sido justamente en los años noventa cuando España ha comenzado a desarrollar estas evaluaciones, con la creación del Instituto Nacional para la Calidad de la Educación, hay razón para dudar de tales afirmaciones.

El artículo de Phelps analiza la situación de 31 países o provincias, y muestra que en la gran mayoría el uso de pruebas en gran escala está aumentando claramente: 27 países o provincias han incrementado el número de pruebas estandarizadas en gran escala, en el cuarto de siglo transcurrido de 1974 a 1999. Esta cifra incluye Alemania, Bélgica, Canadá, China, Dinamarca, Escocia, España, Finlandia, Francia, Hungría, Inglaterra, Irlanda, Japón, Holanda, Nueva Zelanda, Portugal, la Rep. Checa y Suecia y las provincias canadienses de Alberta, Columbia Británica, Manitoba, New Brunswick, Newfoundland, Nueva Escocia, Ontario, Quebec y Saskatchewan. Corea mantuvo la situación, eliminando unos exámenes para substituirlos por otros. En sentido contrario, solamente Australia, Grecia y la provincia canadienses de la Isla del Príncipe Eduardo muestran un decremento en las pruebas en gran escala (Phelps, 2000: 13-15).

Otras fuentes permiten afirmar que la tendencia es la misma en países no cubiertos por el análisis anterior. En América Latina, además de México, Chile destaca por la introducción, desde 1981, del Sistema para la Medición de la Calidad Educativa (SIMCE); pero hay también pruebas en gran escala en Argentina desde 1993; en Brasil desde 1990; en Colombia desde 1990; y en Costa Rica desde 1987 (Wolff, 1998). Honduras

también está desarrollando un sistema de este tipo, y el resto de los países de América Central, lo mismo que Bolivia, implementan sistemas de evaluación para la educación superior, aunque por ahora no incluyan pruebas en gran escala. La Oficina Regional de la UNESCO (la OREALC) comenzó en 1995 a implantar un programa regional de evaluación de la calidad de la educación básica, que ha implicado la aplicación de pruebas en gran escala en los países de la región, cuyos primeros resultados han aparecido ya (Wolff, 1998: 21).

En África sobresale la iniciativa de varios países del sur del continente, que han desarrollado conjuntamente un sistema de evaluación con pruebas en gran escala: el *South African Consortium for the Monitoring of Educational Quality* , SACMEQ.

Los trabajos de evaluación comparada a nivel internacional

Aun habiendo sistemas nacionales de evaluación, la comparación de sus resultados no es sencilla, dadas las diferencias considerables de los sistemas educativos, en cuanto a estructura, curricula y calendarios escolares, además de las diferencias de contenido, grado de dificultad y enfoque de los instrumentos de evaluación mismos. Por ello son importantes estudios comparativos, como los de la *International Association for the Evaluation of Educational Achievement* (IEA).

Ante la insatisfacción por el uso de tasas de graduación o eficiencia terminal de un nivel educativo como único indicador de calidad, un grupo reunido en el Instituto de Educación de Hamburgo planteó en 1958, la posibilidad de un trabajo que diera resultados internacionales estrictamente comparables, con instrumentos equivalentes en contenido y dificultad. El primer estudio piloto, con muestras reducidas de 12 países, se organizó en 1959; los datos se recogieron en 60-61, se procesaron en 61 y se publicaron en 62. Tras esa experiencia se diseñó el primer trabajo en gran escala, sobre matemáticas, en los mismos 12 países, pero con muestras mayores. La etapa central de recolección de datos tuvo lugar en 1964, y se desarrolló con financiamiento norteamericano.

En 1966 la IEA se constituyó formalmente y durante el resto de la década y la siguiente condujo nuevos trabajos: sobre ciencias en 19 países; sobre lectura de comprensión en 15; literatura en diez; educación cívica en diez; francés e inglés como segunda lengua en 18; matemáticas, ciencias e historia con el entorno del aula en diez países. En los ochenta la IEA llevó a cabo un segundo estudio sobre matemáticas, en 20 países; otro sobre ciencias en 24; y otro sobre composición escrita en 14. Del fin de los ochenta a mediados de los noventa la Asociación condujo un trabajo más sobre el uso de computadoras en educación, en 23 países; otro sobre pre-primaria en 14; uno más sobre lectoescritura (*reading literacy*) en 31; y el tercer estudio sobre matemáticas y ciencias, en el que participaron más de 40 países (*Third International Mathematics & Science Study* , TIMSS) (Husén y Postlethwaite, 1996).

Hasta principios de los noventa hubo otros trabajos en la dirección de construir sistemas de evaluación que permitieran análisis comparativos a nivel internacional, en particular los del *Educational Testing Service* conocidos con el nombre *International Assessment of Educational Progress* (IAEP), a partir del programa *National Assessment of Educational Progress* (NAEP). Estos trabajos, sin embargo, no han tenido continuidad, por lo que puede sostenerse que: "En el campo de la evaluación comparativa hay pocas dudas de que, desde sus inicios a principios de los años sesenta, esta organización (la IEA) ha sido la principal fuente de comparaciones confiables entre sistemas educativos" (Goldstein, 1996: 125). A fines de la década aparecieron otras iniciativas internacionales, entre las que pueden mencionarse los trabajos del Observatorio Latinoamericano de la Calidad de la Educación, y los de la Organización para la Cooperación y Desarrollo Económico.

La situación americana reciente

El sistema educativo de Estados Unidos sigue siendo excepcional, por sus dimensiones y por la antigüedad, variedad y solidez de sus sistemas de evaluación. Por ello conviene volver sobre su situación que, según unas opiniones, estaría apartándose de su tradición. Lejos de caracterizarse por el abandono de las pruebas en gran escala, puede afirmarse que la situación actual en los Estados Unidos se distingue por la búsqueda de formas de asegurar la comparabilidad e integración de los resultados de un número creciente de sistemas de evaluación, desarrollados por cada una de sus entidades federativas, con referentes curriculares y criterios heterogéneos.

Desde 1963 se trató de enfrentar este problema mediante un sistema nacional de evaluación que comenzó a concebirse a partir de las ideas de Ralph Tyler y, tras no pocas dificultades, en 1969 se concretó en el *National Assessment of Educational Progress* (NAEP) que aplica, a muestras nacionales y estatales de alumnos, pruebas avanzadas con muestreo matricial de los contenidos y procedimientos controlados, pese a lo cual sus limitaciones son claras: no permiten dar resultados a nivel de plantel, ni siquiera de distrito, y sólo cubren ciertas materias y grados: lectura de 4º grado en 1992, 94 y 98; escritura en 8º grado en 98; matemáticas en 4º grado en 1992 y 96, y en 8º grado en 1990, 92, y 96; ciencia en 8º grado en 1996. En los últimos años el trabajo del NAEP continúa, y la llamada ronda 2000, que está en curso, es la más grande emprendida hasta ahora por el programa. Además de ello, y tratando de superar las limitaciones mencionadas, conjuntando cobertura amplia de contenidos, calidad técnica de las pruebas y posibilidad de analizar resultados a nivel de distrito y plantel, combinación que ni siquiera un programa tan fuerte como el NAEP logra satisfactoriamente, se buscan otras soluciones: la prueba nacional voluntaria (*Voluntary National Test* , VNT), propuesta por el Presidente Clinton en su mensaje sobre el Estado de la Unión de 1997, o un gran sistema de pruebas adaptativas computarizadas, aplicado por Internet, que propone la *Rand Corporation* (Klein y Hamilton, 1999).

La realidad, pues, es que tanto en los Estados Unidos como en muchos otros países, lejos de abandonarse, la aplicación de pruebas en gran escala sigue extendiéndose. Otra cosa es que también las pruebas convencionales estén siendo complementadas y, en algunos casos, substituidas, por pruebas más avanzadas, que utilizan los adelantos teóricos y técnicos descritos antes. También debe distinguirse los caso en que las nuevas pruebas se utilizan para evaluar sistemas y subsistemas educativos con propósitos de monitoreo, planeación y rendimiento de cuentas (*accountability*) y los casos en que se usan para evaluar a los alumnos en lo individual, complementando otros mecanismos basados principalmente en la responsabilidad de los maestros y manejados a nivel de plantel o de distrito escolar (*Cfr.* Martínez Rizo, 1996).

La evaluación y la medición educativa en México

La existencia de instituciones o dependencias dedicadas profesionalmente a elaborar y aplicar instrumentos estandarizados para la evaluación de habilidades académicas y aprendizajes es muy reciente en México. Hay antecedentes relativamente antiguos, pero por alguna razón esos esfuerzos tempranos no lograron consolidarse, y no fue sino hasta la última década del siglo XX cuando se puede hablar de organismos profesionales de elaboración de pruebas.

De 1936 a 1958

El antecedente más remoto de la investigación y la medición educativa en nuestro país es el Instituto Nacional de Psicopedagogía, creado en 1936 por Lázaro Cárdenas. El INP, sin embargo, no desarrolló una actividad significativa en los períodos presidenciales del propio Cárdenas y los de Avila Camacho y Alemán, al grado de que al comenzar el sexenio de Ruiz Cortines estuvo en riesgo de desaparecer (*Cfr.* DGESIC, ca. 1959: 53).

Aunque no es fácil contar con información sobre los inicios del INP, hay indicios de una actividad interesante que, de haber tenido continuidad, podría haber conducido a un desarrollo más temprano de la investigación psicométrica en el país. Una obra de la época informa que, ante la insatisfacción que provocaba el contar solamente con una prueba de tipo verbal para evaluar la inteligencia ("la estandarización Santamarina de la Prueba Individual Binet-Simon"), desde finales de 1936 se decidió llenar la laguna mediante "la estandarización de una prueba de ejecución que permita medir la habilidad mental a través del movimiento, que es otra de sus expresiones funcionales".

El trabajo de estandarización fue hecho por el Prof. Matías López, del Servicio de Psicometría, con ayuda de las Profas. Etelvina Acosta, Guadalupe Díaz y Ma. Eulalia Benavides. El informe, que utiliza una metodología simple, pero cuidadosamente seguida, muestra también un conocimiento de los autores más importantes del campo, ya que se cita pertinentemente a Wundt, Neuman y James, pero también a Bain, Royce y Thorndike, así como a autores menos conocidos, como Hobhouse, Tichtener, Maudsley, Ladd y Woodworth (López *et al.* , 1938: 11-27).

Al comenzar el sexenio 1952-58, la gravedad de la problemática educativa, en un momento en que el crecimiento demográfico comenzaba a volverse explosivo, hacía más necesarios que nunca trabajos rigurosos de investigación educativa, y se decidió apoyar al Instituto en la tarea de desarrollar ese tipo de estudios, en colaboración con el nuevo Consejo Nacional Técnico de la Educación. Se incrementó, pues, el presupuesto y se crearon plazas técnicas y administrativas. Así, el total del personal adscrito aumentó de 68 personas en el sexenio 1946-52 a 95 en 1952-58.

El informe de labores del sexenio 1952-58 reporta que en ese periodo los laboratorios atendieron a 1,800 personas en exámenes clínicos; 4,000 en exámenes médicos; 2,400 en exámenes de agudeza visual y 552 para estudios y tratamientos de ortolalia. Además de hicieron 3,600 estudios socioeconómicos y se aplicaron ¡¡pruebas mentales¿¿ a niños de primaria y secundaria. En este último aspecto se reportan 160,000 pruebas en 1953; 135,000 en 1954; 140,000 en 1955; y 160,000 en 1956. Cada uno de estos años las pruebas se aplicaron a un número de escuelas que varió de 122 a 140; se cubrieron todos los grados, con un número mayor en 1° de primaria (de 20,000 a 26,000) y cifras algo menores (de 10,000 a 20,000) en los demás, hasta 3° de secundaria (DGESIC, ca. 1959: 54-59)

La orientación del modesto trabajo de investigación que se realizaba en el INP puede inferirse de los datos anteriores: una vertiente fundamental era de tipo médico; la otra incluía la aplicación de pruebas, pero sin indicio alguno de que se utilizaran los avances hechos para mediados de siglo por la psicometría norteamericana, pese a que dos décadas antes los trabajos del Prof. Matías López mostraban conocimiento de los trabajos de las primeras décadas del siglo.

De 1958 a 1990

La década del sesenta es la época del inicio real de la investigación educativa en México, con la fundación del Centro de Estudios Educativos, en 1963. Sin embargo, el nuevo Centro atendió otras dimensiones de la problemática educativa, en especial de tipo social; el desarrollo de aspectos psicométricos se dio en la Escuela de Psicología de la UNAM, pero en ella el interés educativo fue también secundario, frente al clínico y otros; ninguna institución cultivó sistemática y profesionalmente la psicometría, y la elaboración de pruebas objetivas de aprendizaje comenzó artesanalmente, con propósitos de selección, en las universidades, en los sesenta y principios de los setenta.

Posiblemente la experiencia más antigua sea la de la Facultad de Medicina de la UNAM, que a mediados de los sesenta desarrolló un banco de reactivos de opción múltiple para el examen de titulación de la carrera de Médico Cirujano, calificado por computadora. Fue importante también la experiencia de la Facultad de Ingeniería donde, desde 1976, y además de pruebas de diagnóstico para alumnos de nuevo ingreso, se incursionó en forma imaginativa en evaluación, con exámenes computarizados de matemáticas cuyos resultados se daban a los maestros, el coordinador de la carrera, el director de la división y la dirección de la Facultad, además de las escuelas de procedencia de los aspirantes; se desarrolló un banco de reactivos como apoyo a los profesores para que mejoraran sus propias evaluaciones; se desarrolló un programa de calificación que permitía manejar reactivos de opción múltiple, de falso verdadero, jerarquización, relación de columnas e

inclusive de complementación y respuesta abierta breve (Información personal del Ing. Agustín Tristán).

Sin embargo, seguramente la experiencia más importante fue la relativa a los exámenes de selección para ingreso a la UNAM, tanto en el nivel de bachillerato como en el de licenciatura, procesos que se volvieron complejos y delicados cuando el número de aspirantes se multiplicó y los lugares disponibles se volvieron cada año más insuficientes.

Para atender esta problemática, según informa el Mtro. Rafael Vidal, Director Técnico del CENEVAL, la UNAM estableció en 1963 un área especializada a cuyo frente estuvo durante muchos años la Dra. Emma Dolujanof, quien sin duda debe ser considerada la persona más importante en este terreno en la etapa a que se refiere este apartado. Sin un acercamiento sofisticado, pero con gran cuidado y competencia, la Dra. Dolujanof logró desarrollar pruebas que, sometidas *a posteriori* a evaluaciones psicométricas, mostraron niveles de confiabilidad excelentes e, inclusive, grados de equivalencia de la versión de un año a las siguientes no muy inferiores a los que se alcanzan con las modernas técnicas de igualación o equiparación (*equating*). Los trabajos de esta área de la UNAM, continuados luego en el marco de la Dirección General de Planeación, fueron la base para la creación del CENEVAL en 1994.

Otras IES privadas (como la Universidad Iberoamericana) o públicas (como la Universidad Autónoma de Aguascalientes) desarrollaron sus propias pruebas de ingreso durante los años setenta, también sin controlar sistemáticamente las propiedades psicométricas de las pruebas y, en consecuencia, sin poder hacer análisis comparativos o longitudinales. Otras instituciones privadas, como el ITESM y la Universidad de las Américas, comenzaron a utilizar los servicios de la Oficina establecida por el *College Board* en Puerto Rico en 1963, contratando la aplicación de la Prueba de Aptitud Académica (PAA), equivalente al SAT, para propósitos de selección para el ingreso al nivel de licenciatura.

En lo que hace a educación básica, desde fines de los setenta la SEP creó un área de evaluación para valorar el aprendizaje de muestras nacionales de alumnos de primaria. También en este caso, sin embargo, el trabajo era "artesanal": no se dominaban las técnicas psicométricas plenamente establecidas en esas fechas; no se controlaba sistemáticamente la confiabilidad y la validez de los instrumentos; no se verificaba la equiparabilidad de las versiones aplicadas en años sucesivos ni la estabilidad de las escalas de medición utilizadas.

Los últimos años

Para ver nacer organismos dedicados a la elaboración de pruebas que atiendan en forma sistemática cuestiones técnicas deberá esperarse hasta los años noventa, con la fundación del CENEVAL (1994) y la modernización del área de evaluación de la SEP (1995), precedidas por el lanzamiento del Examen de Conocimientos y Habilidades Básicas (EXCOHBA) de la Universidad Autónoma de Baja California (1990). De la misma fecha data la incursión masiva en el mercado mexicano de las pruebas de aptitud académica en español de la oficina de Puerto Rico del *College Board* .

En términos cuantitativos el número de pruebas aplicadas anualmente se incrementó en forma espectacular durante la segunda mitad de la década de 1990: el CENEVAL pasó de aplicar 365,552 pruebas en el ciclo 1994-95 a más de un millón en el año 2000. El *College Board* , que en 1989 aplicaba anualmente unas 25,000 Pruebas de Aptitud en toda América Latina, en 1999 aplicaba ya más de 50,000 sólo en México. La SEP, además de las pruebas tradicionales a muestras nacionales aplicadas desde fines de los setenta, a mediados de los noventa comenzó a aplicar unos cinco millones de pruebas anuales para Carrera Magisterial y medio millón más de las pruebas de estándares curriculares. A los anteriores esfuerzos nacionales se añadieron, también en los noventa, los de algunas entidades que comenzaron a implantar sus propios sistemas de evaluación, como el estado de Aguascalientes.

El principal avance, en términos cualitativos, es la aparición en el país de instancias profesionales de elaboración de pruebas (CENEVAL, SEP) que producen instrumentos que se acercan lentamente a estándares de calidad internacionales. Hay esfuerzos institucionales dignos de mención, como los de la UABC, con una versión de aplicación por computadora, no adaptativa, del EXCOHBA, y que trabaja en pruebas adaptativas; los del ITESM, que ofrece a sus alumnos exámenes personalizados por computadora en diversas materias,

los del sistema de universidades tecnológicas en el mismo sentido, entre otros.

Sin embargo, la falta de tradición psicométrica hace que todavía haya limitantes serias, no todas las cuales pueden imputarse a los responsables de las pruebas. De parte de éstos, ciertamente, debe señalarse que los avances más importantes de la psicometría de la segunda mitad del siglo XX siguen siendo casi desconocidos. Las mejoras técnicas se limitan a aplicar, con limitaciones, la teoría clásica, en tanto que la de la generalización y las variantes de la de respuesta al item son casi totalmente ignoradas. Las pruebas siguen siendo de opción múltiple, en tanto que las de respuesta construida y ejecución no existen, lo mismo que pruebas adaptativas, o de papel y lápiz con muestreo matricial para aplicaciones en gran escala.

A más de esto, hay limitaciones importantes que son responsabilidad de otras instancias: no pocas IES, por ejemplo, pese a las recomendaciones técnicas, utilizan los resultados en las pruebas estandarizadas de ingreso como único elemento de decisión, desoyendo las recomendaciones técnicas de incorporar otros elementos, como el promedio del bachillerato. Otra limitación no técnica muy importante es el manejo de los resultados de las evaluaciones en gran escala, que algunas autoridades parecen considerar secreto de estado, en lugar de darles la debida difusión, con el debido cuidado de la privacidad de las personas involucradas, difusión que es fundamental para los propósitos de mejoramiento, en relación con el concepto de rendimiento de cuentas (*accountability*), difusión que no se hace seguramente por motivos políticos.

Discusión y conclusiones

La evaluación no es, en sí misma, un fin. Emplear instrumentos objetivos de medición del aprendizaje que alcanzan los alumnos de una institución o sistema educativo tendrá valor en la medida en que los resultados se aprovechen para el mejoramiento de tales instituciones y sistemas.

Para sustentar la afirmación anterior, comenzamos aceptando sin dificultad que las pruebas objetivas no pueden substituir al maestro en lo que se refiere a la evaluación del aprendizaje de cada alumno en lo individual. Los docentes, sobre todo si trabajan como equipo en cada escuela, deberán seguir siendo, junto con los propios alumnos, los actores fundamentales tanto del binomio aprendizaje-enseñanza como de la evaluación de sus dos componentes.

Con igual claridad se sostiene que las evaluaciones que hacen los maestros de sus propios alumnos, por su misma naturaleza contextualizada y personalizada, no pueden ser estrictamente comparables entre sí, y no son suficientes para valorar el nivel de aprendizaje que alcanzan en conjunto los millones de alumnos que comprenden los sistemas educativos de muchos países, o los millares de muchas IES. Para estas valoraciones de conjunto, necesarias para el establecimiento de políticas públicas en el campo educativo, es indispensable usar instrumentos que reúnan tres tipos de características que las evaluaciones individuales no pueden conjuntar:

- Las cualidades psicométricas básicas de validez en sus variantes (que lo que mide el instrumento corresponda a los objetivos y contenidos del currículo), confiabilidad (que las mediciones sean consistentes) y ausencia de sesgo;
- Otras cualidades técnicas que permitan comparaciones tanto transversales (entre diversas partes del sistema educativo -como entidades federativas- o con otros países) como longitudinales (esto es a lo largo del tiempo): equiparabilidad de versiones, estabilidad de la escala;
- Otras características que permitan que los instrumentos puedan aplicarse a muestras numerosas, compuestas por decenas de miles de sujetos, en forma casi simultánea, y que los resultados sean procesados rápida y confiablemente, todo ello dentro de límites razonables de costo.

Los instrumentos que reúnen estos tres tipos de características son, precisamente, las llamadas *pruebas de aplicación en gran escala* , que frecuentemente se identifican con exámenes de papel y lápiz formados exclusivamente por preguntas de opción múltiple. Cuando estas pruebas se elaboran deficientemente dan pie a las críticas que las señalan como superficiales y, en el peor de los casos, sesgadas social y culturalmente.

Si se aprovechan los desarrollos metodológicos y técnicos actuales, en cambio, es posible contar no sólo con pruebas tradicionales de buena calidad, sino con instrumentos sofisticados muy finos, incluyendo pruebas de respuesta construida y de ejecución, aplicadas por computadora, adaptativas, etc. La disyuntiva que opone aplicación en gran escala a calidad y finura del instrumento es, pues, falsa.

Dichos avances permiten, pues, en principio, que las autoridades cuenten con instrumentos adecuados para hacer las evaluaciones masivas y finas que se requieren para sustentar las políticas de mejoramiento del sistema educativo mexicano y de las IES nacionales. Para que esta posibilidad se vuelva realidad, sin embargo, y dada la complejidad técnica del campo, se necesita que se consolide la comunidad de especialistas en medición y evaluación educativa. Si esto no sucede los instrumentos disponibles seguirán careciendo del nivel necesario, o el país dependerá, también en este campo, del mercado internacional, con obvias desventajas de costo, riesgo de inadecuación curricular y cultural, y dependencia tecnológica.

Para evitar tal dependencia, y dado el desfase que separa los avances en Estados Unidos y otros países de los equivalentes en el nuestro, es necesario analizar la situación de quienes han incursionado antes en este tipo de actividades, no para imitarlos ciegamente, pero sí para aprovechar su experiencia en forma crítica. Esto resulta necesario porque, al desarrollarse un campo y formarse masas críticas de especialistas, se generan procesos acumulativos que no ocurren en contextos de menor desarrollo, de manera que el avance dificultoso de las primeras etapas puede ir seguido por otros más rápidos. El conocimiento de experiencias tempranas podrá ahorrar costosos errores y permitirá avanzar con rapidez creciente para acercarse a los punteros de la carrera; el aislamiento, por su parte, significará también que el rezago se vaya haciendo cada vez mayor. Tal estrategia es tanto más necesaria cuanto que parece haber una notable similitud de las situaciones que se comparan, a la distancia de algunas décadas, como se muestra para terminar.

Con las excepciones señaladas en el trabajo, en México las pruebas de opción múltiple de buena calidad técnica comienzan apenas a extenderse y se considera novedosa la utilización de la teoría clásica de las pruebas en su diseño y la interpretación de sus resultados. Esto corresponde, como hemos visto, a lo que pasó en los Estados Unidos entre 1930 y 1960; en México, sin embargo, tiene lugar enfrentando resistencias y rechazos que en el país del norte se presentaron hasta los setenta y los ochenta; se hace, además, en un contexto en el que tanto los partidarios de las pruebas como sus críticos parecen desconocer los avances que, a partir de los sesenta, y en forma muy amplia en los noventa, permiten corregir de manera bastante satisfactoria muchas de las limitaciones reales de las pruebas tradicionales, atendiendo así las inquietudes válidas que, en parte, explican las resistencias y rechazos mencionados.

El paralelismo de algunos rasgos de la historia del *College Board* , a los que se ha hecho referencia antes, y la del CENEVAL es, también, notable. En el Prefacio de la edición de 1984 del *Technical Handbook for the SAT* se subraya, por ejemplo, que la historia del organismo "no es la de una agencia productora de pruebas que trata de vender sus servicios a las universidades, sino la de una asociación de *colleges* que se unen con el propósito común de facilitar el paso de los alumnos de la *high school* a la universidad (Donlon, 1984: xix).

Al presentar algunos aspectos de la historia norteamericana y mundial de las evaluaciones en gran escala, este artículo ha mostrado que los problemas que enfrentan las instituciones que se dedican a esa tarea en nuestro país no son inéditos, sino que se han presentado en otros lugares, que hay soluciones bien probadas ya para algunos, y que en el último medio siglo los avances en este campo han sido muy importantes, aunque haya todavía muchos retos por atender.

Si la naciente comunidad de especialistas mexicanos en temas de medición y evaluación educativa aprovecha la historia de las pruebas en gran escala, y la literatura que refiere la manera en que teoría y técnicas avanzan para enfrentar los problemas detectados en la práctica, el trabajo de producir buenos instrumentos de evaluación, para el gran propósito de contribuir al mejoramiento de las escuelas y los alumnos del país, se verá grandemente beneficiado.

Referencias

AERA-APA-NCME (1999). *Standards for educational and psychological testing* . Washington, American Educational Research Association.

BRIGHAM, Carl C. et al. (1926). "The Scholastic Aptitude Test of the College Entrance Examination Board", en Fiske, Thomas S. (Ed.). *The Work of the College Entrance Examination Boardm* , 1901-1925. New York, Ginn & Co. pp. 44-63. Citado en Donlon (1984: 2).

CENEVAL (1999). *Informe de resultados 1998* . México, CENEVAL.

CRONBACH, L. J., G. C. Gleser, H. Nanda y N. Rajaratnam (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles* . New York, Wiley.

DIR. GRAL. DE ENSEÑANZA SUPERIOR E INV. CIENTÍFICA (ca. 1959). *Enseñanza Superior e Investigación Científica. Seis años de labor 1952-1958* . México, SEP.

DONLON, Thomas F. (Ed.) (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests* . New York, College Entrance Examination Board. Especialmente los capítulos I (The Admissions Testing Program: A Historical Overview, pp. 1-11) y II (Psychometric Methods Used in the Admissions Testing Program, pp. 13-36).

FREDERIKSEN, Norman, Robert J. Mislevy e Isaac I. Bejar (Eds.) (1993). *Test Theory for a New Generation of Tests* . Hillsdale, New Jersey, Lawrence Erlbaum.

GOLDSTEIN, Harvey (1996). "Introduction". *Assessment in Education: principles, policy & practice* . Vol. 3 (July) No. 2, pp. 125-128.

HUSÉN, Torsten y T. Neville Postlethwaite (1996). "A Brief History of the International Association for the Evaluation of Educational Achievement (IEA)". *Assessment in Education: principles, policy & practice* . Vol. 3 (July) No. 2, pp. 129-141.

KEEVES, J. P. (1997). "Section III. Measurement in Educational Research. Introduction: Advances in Measurement in Education", en Keeves, John P. (Ed.) *Educational Research, Methodology, and Measurement. An International Handbook* . Oxford-New York-Tokyo, Pergamon, pp. 705-712.

KLEIN, Stephen P. y Laura Hamilton (1999). *Large-Scale Testing. Current Practices and New Directions* . Santa Monica, Rand Education.

LANDSHEERE, Gilbert de (1996). *La investigación educativa en el mundo* . México, FCE.

LÓPEZ, MATÍAS et al. (1938). *Medida de la inteligencia. Prueba individual de ejecución de "Kohs"* . México, SEP.

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

MARTÍNEZ RIZO, Felipe (1996). *La calidad de la educación en Aguascalientes. Diseño de un sistema de monitoreo* . Aguascalientes. UAA-IEA.

MATEO ANDRÉS, Joan (1999). *Enciclopedia General de la Educación* . Barcelona, Océano. Vol. 2, Sección VI, Evaluación e investigación, pp. 529-676.

MISLEVY, Robert J. (1993). "Foundations of a New Test Theory", en Frederiksen et al. 1993, pp. 19-39.

PHELPS, Richard P. (2000). "Trends in Large-Scale Testing Outside the United States". *Educational Measurement: Issues and Practice* (Spring), pp. 11-21.

WOLFF, Lawrence (1998). *Las evaluaciones educacionales en América Latina: Avance actual y futuros desafíos* . Santiago de Chile, PREAL.