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

Grupo de Trabajo sobre Estándares y Evaluación

Hecho a la Medida

David J. Hoff

Grupo de Análisis para el Desarrollo

HECHO A LA MEDIDA

David J. Hoff

*Tomado y traducido de la sección especial de **Education Week** del 16 de junio de 1999 dedicada a la cultura evaluativa (pp. 21 a 27), con autorización de sus editores.*

Las pruebas estandarizadas son herramientas de los gobiernos estatales y federales que orientan qué es lo que se enseña y cómo se está enseñando.

Al llegar cada primavera, los estudiantes del tercer al décimo grados de Tejas rinden la Evaluación de Habilidades Académicas de Tejas (*Texas Assessment of Academic Skills*).

Ellos trabajan por lo menos dos días completos rellenando los círculos en que se marcan las respuestas a ítems de opción múltiple, así como preguntas que requieren como respuesta un "verdadero" o "falso", en matemáticas y lectura. Los de octavo grado dedican otros dos días a pruebas de estudios sociales y ciencias. Para graduarse, los alumnos de secundaria deben aprobar tales pruebas en lectura, escritura y matemáticas.

También están en juego los derechos a vanagloriarse de sus escuelas, la reputación de sus maestros e incluso, los valores de sus hogares.

Los programas de pruebas como el de Tejas se están volviendo comunes en los estados al acercarse el final del siglo. Cuarentiocho estados tienen sistemas de pruebas y la mayoría se apoya en sus resultados para determinar una amplia variedad de acontecimientos de impacto vital, incluyendo si los maestros recibirán bonificaciones o si la escuela llevará una divisa de honor -- o de vergüenza -- debido a sus puntajes. Tejas y otros 18 estados también exigen a los estudiantes aprobar un examen para obtener un diploma.

Esta generación actual de pruebas es la culminación de un siglo en el cual las pruebas y otras evaluaciones -- que van desde pruebas de coeficiente intelectual, pasando por la prueba de aptitud académica (Scholastic Aptitude Test – SAT), hasta pruebas de proficiencia a nivel estatal -- vienen jugando un papel cada vez más omnipresente en las vidas de los estudiantes y en el funcionamiento de las escuelas a las que asisten.

A comienzos del siglo XX, las pruebas estandarizadas eran potestad de los distritos urbanos y eran empleadas mayormente para medir qué tan bien estaban aprendiendo los estudiantes. Usualmente, los puntajes de las pruebas sólo eran accesibles a un número selecto de administradores.

Cien años más tarde, los exámenes son instrumentos de los gobiernos estatales y el federal, dirigiendo o influenciando lo que se enseña y cómo es enseñado.

La creencia en el valor de la medición de logros está enraizada en los ideales políticos que dieron forma a los E.E.U.U, argumenta Eva L. Baker, co-directora del Centro de Investigación sobre Estándares y Evaluación de Estudiantes, un proyecto de

investigación financiado por el gobierno federal, y profesora en la Universidad de California de Los Angeles.

Según Baker, Alexis de Tocqueville, el astuto observador de los Estados Unidos a comienzos del siglo XIX, notó que los norteamericanos creían en la "perfectabilidad del hombre", y la obsesión norteamericana por administrar pruebas en el siglo XX refleja ese credo. Los educadores continúan buscando el instrumento perfecto que los ayude a proporcionar el mejor programa educativo posible para sus estudiantes.

Pero el uso de pruebas también levanta numerosas cuestiones acerca de otro preciado principio norteamericano: la igualdad de oportunidades.

Desde los movimientos por los derechos civiles y los derechos de las mujeres en los sesentas y setentas, las escuelas y universidades se han visto bombardeadas con quejas y litigios de críticos que sostienen que las pruebas estandarizadas no son justas con los estudiantes de las minorías ni con las mujeres. Los sesgos de las pruebas producen puntajes que encarrilan a los estudiantes hacía cursos de educación especial y hacia cursos menos desafiantes que los apartan de la posibilidad de ingresar a las universidades competitivas.

Mientras tanto, los líderes escolares se quejan diciendo que las pruebas son barómetros injustos de qué tan bien están educando a los chicos y que no deberían ser el único criterio usado para tomar decisiones sobre el futuro de los estudiantes, tales como si un alumno de décimo grado de Brownsville, Tejas o Newton, Massachusetts se graduará de la secundaria.

El Examen Impreso

Ya en 1845, las escuelas empezaron a aplicar pruebas a sus estudiantes de manera uniforme. Por entonces, los exámenes orales eran la forma de evaluación por excelencia.

Según **Cómo la investigación cambió a las escuelas norteamericanas**, escrito por Robert M. W. Travers, Boston se convirtió ese año en el primer distrito norteamericano en imprimir pruebas de respuestas cortas que se usarían en todo su sistema. Los estudiantes eran evaluados en geografía, gramática, historia, retórica y filosofía. El Consejo Escolar de la ciudad propuso aplicar las pruebas a todos los estudiantes de sus escuelas, pero sólo unos 20 o 30 estudiantes de cada una las rindieron. De los 7000 estudiantes que tenía la ciudad, sólo unos quinientos tomaron cada una de las pruebas.

Travers escribe en su libro de 1983 que, en una reacción que sonará familiar a los educadores de hoy en día, los líderes del sistema escolar de Boston quedaron impactados por el pobre desempeño. Los alumnos examinados fallaron en responder correctamente el 40 por ciento de las preguntas en cualquiera de los temas. Los resultados, según un reporte contemporáneo sobre las pruebas, "mostraron más allá de toda duda, que una gran proporción de los escolares de nuestras primeras clases, niños y niñas de 14 ó 15 años de edad, cuando se les pedía que escribieran oraciones sencillas para expresar sus pensamientos sobre ciertos temas comunes, no podían escribir sin errores de gramática, ortografía o puntuación sin la ayuda de un diccionario o un profesor."

Según Travers, las pruebas se volvieron a dar en 1846, pero ya en 1850 Boston había abandonado su estrategia, retornando hacia las pruebas no estandarizadas, mayormente basadas en presentaciones orales.

Según David B. Tyack en **El Mejor Sistema: Una Historia acerca de la Educación Urbana Norteamericana**, Portland, Maine, experimentó por primera vez con pruebas estandarizadas en 1874. El superintendente de Portland, Samuel King, creó un currículo único de estudios para las escuelas de la ciudad y escribió una prueba para medir si los estudiantes lo aprendían de manera exitosa.

Según documentos citados por Tyack en su obra escrita en 1974, el superintendente escribió acerca de este examen: "Sistema, orden, prontitud y puntualidad han caracterizado los exámenes y ejercido una influencia útil sobre los alumnos, al estimularlos a estar concienzudamente preparados para cumplir sus compromisos y citas".

King publicó el puntaje de cada estudiante en el periódico, despertando la ira y la oposición de maestros y padres, lo que llevó a su renuncia en 1877. Su sucesor nunca publicó los puntajes pero continuó con las pruebas, aunque las hizo más fáciles con el fin de lograr que las aprobaran más estudiantes que los que aprobaron durante la gestión de King.

La Emergencia del Coeficiente Intelectual

Las experiencias de Boston y Portland fueron comunes en distritos urbanos hasta principios del siguiente siglo. Pero en los primeros dos decenios del siglo veinte, investigadores educativos como Edward L. Thorndike empezaron a crear pruebas estandarizadas para medir a los estudiantes en escalas uniformes, tanto en aritmética como en caligrafía y otras asignaturas.

Las pruebas estandarizadas tenían un gran atractivo porque eliminaban la subjetividad de los métodos individuales de calificación de cada profesor. Los expertos afirmaban que con una prueba objetiva el puntaje de un alumno podría compararse con confianza con el de su compañero de aula y con los de sus contrapartes en una ciudad al otro lado del país.

Todas esas pruebas tenían una meta exclusiva: medir qué tan bien se desempeñaban los alumnos frente a un currículo recomendado. Pero en los primeros años de este siglo, los psicólogos empezaron a preparar una nueva forma de evaluación: una diseñada para medir la habilidad innata y predecir el desempeño futuro, en lugar de evaluar si los estudiantes habían dominado el material de un currículo.

En 1904, el sistema escolar de París contrató a Alfred Binet para diseñar una prueba que identificara a los estudiantes que habían sido incapaces de beneficiarse de la instrucción. El ideó una escala que predecía qué tan bien podría aprender un niño e hizo una estimación de su "edad mental". Los psicólogos norteamericanos Henry Goddard, Lewis M. Terman y otros adaptaron el trabajo de Binet en 1912 para crear el Coeficiente Intelectual, o C.I., calculado dividiendo la "edad mental" de una persona por su edad cronológica.

Terman, un psicólogo de la Universidad de Stanford, dio a conocer la llamada Escala Stanford-Binet en su libro **La medición de la inteligencia**. Allí esbozó cómo debería ser administrada la prueba Binet y explicó cómo sus resultados rendirían datos reveladores sobre la inteligencia innata de un estudiante.

En un desarrollo crucial, la nueva versión de Terman podía ser administrada a los estudiantes usando lápiz y papel, logrando así que las escuelas la administraran de manera mucho más fácil y económica que con versiones anteriores, las cuales requerían los servicios de un psicólogo especialmente entrenando.

Cuando Terman publicó el tercer volumen de la prueba, afirmó que el puntaje de un estudiante sería constante en el curso del tiempo. En 1922, Terman declaró que 250.000 estudiantes habían rendido la prueba.

Harlan C. Hines, profesor de la Universidad de Washington, escribió en la edición de abril de 1922 de la **Revista del Consejo Escolar Norteamericano** que las pruebas de C.I. *"están siendo utilizadas actualmente para clasificar escolares de todos los niveles de inteligencia y están siendo desarrolladas como ayudas importantes para la orientación vocacional"*. En por lo menos un estado, el cual Hines no mencionó, *"los alumnos de secundaria están siendo clasificados únicamente sobre la base de pruebas de inteligencia"*, añadió.

Encarrilamiento[1] *(Tracking)* en base a pruebas

Robert Glaser y Edward Silver, investigadores de la Universidad de Pittsburgh, dicen que, a pesar de que esto constituía un abuso descarado de las pruebas, el encarrilamiento era una práctica común.

A principios de siglo se incrementó la matrícula en las escuelas de la nación, gracias al influjo de inmigrantes y a las leyes de asistencia obligatoria. Los administradores deseaban tener una manera de separar a los alumnos por habilidades debido a que creían que ésta era la mejor forma de otorgarles una instrucción apropiada.

Glaser y Silver escribieron en su veinteava edición de **Revisión de Investigaciones sobre Educación** que las pruebas de inteligencia proporcionaban los datos científicos necesarios para encarrilar a los estudiantes en rutas [académicas] diferenciadas. Aplicar pruebas era *"un instrumento conveniente y poderoso de control social"*, escribieron, y así ha permanecido a lo largo del siglo.

La adopción extendida de la práctica del encarrilamiento de estudiantes reflejaba el consenso de los científicos sociales de esa época respecto a que la inteligencia era un rasgo hereditario. Por entonces se creía que utilizando un instrumento científico para agrupar a los alumnos, los funcionarios escolares podrían encontrar el lugar correcto para atender a los alumnos según sus habilidades innatas.

Hacia mediados de siglo, la Prueba de Aptitud Académica (Scholastic Aptitude Test) se convirtió en un rito anual para los postulantes a universidades y, hacia 1970, los

[1] El encarrilamiento o tracking se refiere a la práctica de agrupar y colocar a estudiantes de un mismo grado en distintas "rutas curriculares", según sus capacidades aparentes. N. del E.

alumnos eran clasificados para educación especial y remedial en programas federales sobre la base de sus puntajes pruebas.

La pregunta tanto en 1920 como ahora, era si aquellas pruebas deberían ser usadas con semejantes propósitos. Basándose en los resultados de las primeras versiones de las pruebas, se decía que los inmigrantes recientes del sur de Europa eran menos inteligentes que otros. Los hallazgos se utilizaron para establecer cupos que restringieran la inmigración desde Italia, Grecia y otros países mediterráneos. Más adelante, un número desproporcionado de estudiantes negros que rendían pobremente en dichas pruebas terminaron en programas de educación especial, en los cuales no llegaban a alcanzar el mismo nivel que sus pares.

Según Glaser, Silver y otros expertos en evaluación, esos usos eran parte de un patrón de uso inapropiado e injusto de tales pruebas. Hines, de la Universidad de Washington, reconoció tempranamente el potencial de problemas que ello acarreaba. Hines concluyó en su artículo de 1922: " El escritor ha llegado a sentir que una prueba pierde su valor y se convierte en un arma peligrosa en manos de personas no entrenadas".

Al año siguiente, Terman elaboró una prueba muy diferente a sus pruebas de inteligencia. Su producto era similar a las pruebas estandarizadas de Boston y Portland del siglo diecinueve, así como a otras que empezaron a llegar a los mercados nacionales en los primeros decenios de este siglo.

En su Prueba de Logros Académicos de Stanford *(Stanford Achievement Test)*, Terman se proponía medir los logros en materias específicas en varios grados. Con ella, un colegio podría medir qué tan bien se desempeñaban sus alumnos de segundo al octavo grados. Más importante aún, los resultados de Stanford podrían ser comparados con aquéllos de una muestra nacional de 350.000 estudiantes, según una biografía escrita en 1988 por Henry L. Minton (**Lewis Terman: Pionero en Pruebas Psicológicas**). La muestra era, de lejos, más amplia que cualquiera de las de pruebas similares de entonces.

Según Glaser y Silver, la forma y contenidos del examen de Stanford "preconfiguraban el futuro". A lo largo del resto del siglo, las pruebas estandarizadas seguirían el modelo del trabajo de Terman.

Midiendo el dominio de contenidos

Una de estas era la Prueba de Destrezas Básicas de Iowa (Iowa Tests of Basic Skills – ITBS), una de las tres pruebas estandarizadas más utilizadas de manera rutinaria actualmente en las escuelas.

El programa de pruebas empezó en 1929 como una competencia por becas de la Universidad de Iowa. Según **Los Programas de Evaluación de Iowa**, una historia del programa escrito por Julia J. Peterson, en el primer año, las pruebas tenían secciones dedicadas a gramática, literatura inglesa y norteamericana, historia universal, historia estadounidense, álgebra, geometría, ciencia, física, mecanografía y taquigrafía.

Los exámenes de Iowa, como muchas pruebas de la época, se basaban en libros de texto comúnmente utilizados en todo el estado. Para acelerar la calificación, los

autores utilizaban preguntas de opción múltiple, “verdadero o falso”, apareamiento y rellenado de espacios en blanco; las únicas preguntas de respuesta libre se presentaban en los exámenes de matemáticas. Las pruebas se podían calificar tan rápidamente que los alumnos sabían sus puntajes y si habían avanzado a la siguiente ronda de la competencia antes de abandonar el auditorio.

Hacia 1935, el programa de becas había sido reemplazado por una batería de pruebas, también llamada Pruebas de Destrezas Básicas de Iowa. Así como la Prueba de Logros Académicos de Stanford, la ITBS cubría todos los grados y abarcaba la currícula básica: lectura, lenguaje, habilidades para el estudio y aritmética. Cada distrito recibía un reporte confidencial clasificando a sus estudiantes con respecto al promedio de una muestra de ámbito estatal, así como el número de sus respuestas correctas. Peterson escribe en su libro de 1983 que el diseñador del examen de Iowa, E.F. Lindquist, pretendía que éste fuera de carácter diagnóstico y que fuera capaz de evaluar si la currícula de un distrito estaba funcionando bien.

A los cuatro años del inicio del programa, 30.000 estudiantes de Iowa habían rendido las pruebas de destrezas básicas. Las pruebas empezaron a proliferar por todo el país, cuando Kansas City, Missouri y Carolina del Sur las adquirieron para utilizarlas en sus aulas.

El mercado para las pruebas continuó ampliándose. En 1940, el departamento de pruebas de la Universidad de Iowa contrató a la compañía Houghton Mifflin de Boston para distribuir su prueba a nivel nacional. Las primeros cuatro versiones aportaron $330.000 en regalías para la universidad.

Las pruebas de Iowa y Stanford no eran las únicas pruebas de logros de la época. Durante los treintas, las de competidores como el Departamento de Pruebas de California también crecieron rápidamente. Travers, en su historia sobre la investigación educativa, escribió: “Los editores habían llegado a reconocer que las pruebas proveían un mercado nuevo y lucrativo que se desarrolló rápidamente en el decenio de 1930, no sólo en lectura, escritura y matemáticas, sino en las áreas de conocimientos de la currícula”.

Una Validación

El decenio de 1940 vio incrementarse el uso de otra prueba estandarizada: la Prueba de Aptitud Académica. Al igual que las pruebas de C.I., el objetivo del SAT entonces y ahora es predecir el desempeño. En lugar de ser un dispositivo para clasificar alumnos dentro de las escuelas, el propósito del SAT era separarlos entre colleges y universidades.

La versión estandarizada del SAT nació en el decenio de 1920, cuando la universidad de Harvard decidió ofrecer becas a estudiantes pobres y a otros que no habían asistido a las escuelas preparatorias de la élite en Nueva Inglaterra. Harvard contrató al Consejo de Evaluación para el Ingreso a las Universidades *(College Entrance Examination Board)* para crear una nueva prueba con el fin de seleccionar a los receptores de dichas becas. Desde 1900, el Consejo, ahora conocido simplemente como el Consejo Universitario *(College Board)*, había ofrecido exámenes de ensayo en campos tales como retórica, Griego y otros elementos del currículo tradicional de las escuelas preparatorias. Aunque el Consejo Universitario continuó ofreciendo los

exámenes de ensayo hasta diez años después de que se diera el primer SAT en 1926, el nuevo examen pronto se convirtió en la valla común que los aspirantes a la universidad necesitaban sobrepasar.

La importancia del SAT se incrementó, junto con las matrículas, cuando los militares regresaron de la Segunda Guerra Mundial y empezaron a acudir en tropel a las universidades al amparo de la ley para los veteranos de guerra.

Brian P. O'Reilly, el director ejecutivo de los programas de orientación vocacional y pruebas de admisión del Consejo Universitario, dice: *"El retorno de los soldados tras la 2a Guerra Mundial fue lo que verdaderamente impulsó el crecimiento de la matrícula en las universidades y por ende, de las pruebas de admisión"*. Dice O'Really que, cuando el SAT estableció en 1941 su escala de 800 puntos para cada una de sus dos secciones, 10.000 estudiantes rindieron la prueba. Siete años más tarde, el número se había duplicado. En 1964, más de un millón la rindieron; ese número se elevó a dos millones en 1967.

El SAT se convirtió en el método preferido para evaluar a los postulantes a la universidad debido a que su realizador, el Servicio de Evaluación Educativa *(Educational Testing Service)*, con ayuda de la IBM, usaba máquinas para revisar rápidamente las hojas de respuestas. Cierto es que los candidatos al programa de becas de Iowa podían saber sus puntajes en unas pocas horas. Pero el número de estudiantes que aplicaban a aquel programa era mínimo en comparación a la invasión de postulantes universitarios de la post-guerra.

Baker, junto a su colega de UCLA Regie Stites, escribieron en un ensayo de 1991 sobre tendencias en las pruebas, que la importancia del SAT radica en que éste "legitimó" las pruebas de opción múltiple, sobre todo entre la clase de personas más educadas que más adelante se convertirían en diseñadores de políticas. *"Nuestros mejores y más brillantes estudiantes y sus influyentes padres, aceptaron la validez de dichas pruebas de admisión a la universidad. Así, la experiencia de haber sido evaluados exitosamente ellos mismos engendró no desprecio sino una reafirmación de la precisión de la medida a ser utilizada con otros."*

Pero los críticos han acusado por largo tiempo al SAT de no ser una medida justa para el resto de la sociedad. Monty Neill, director ejecutivo del Centro para una Evaluación Justa y Transparente, conocido como "FairTest" dice que los datos son claros respecto a que el SAT discrimina a las minorías y a veces a las mujeres. "Esto no ha cambiado sustancialmente (desde sus inicios)". FairTest fue fundado en el decenio de los ochentas por defensores de los derechos civiles y de los consumidores para monitorear las prácticas evaluativas y hacer campaña en contra de aquéllas que sus líderes pensaban eran injustas.

El mayor problema creado por el sesgo puede no radicar en el proceso de admisión, sugiere Neill, sino en las decisiones de los alumnos acerca de adónde aplicar. Un estudiante cuya puntaje combinado en la prueba es de 1000 se sentirá intimidado respecto a la posibilidad de solicitar admisión a universidades que publican SATs promedio de 1100, dice. *"Refuerza las jerarquías de clase y raza existentes"*, asegura.

El Consejo Universitario sostiene que el SAT es el segundo mejor predictor de qué también se desempeñará un alumno en la universidad.

O'Reilly dice: "Nos sentimos cómodos sabiendo que las calificaciones obtenidas en la secundaria son los mejores predictores, pero sólo muy poco por encima de la puntuación obtenida en las pruebas. Ninguna otra cosa contribuye mucho".

Las calificaciones obtenidas en la secundaria se derivan mayormente de las pruebas preparadas por los profesores -- las pruebas "que cuentan" ante los ojos de los alumnos. Hay poca investigación disponible que muestre cómo han cambiado estas evaluaciones a lo largo del siglo.

El SAT no es, añade O'Reilly, un buen barómetro de la calidad de la escuela.

Desde el decenio de los setenta, agentes de bienes raíces, editores de periódicos y secretarios de educación norteamericanos han citado datos del SAT en sus intentos de evaluar la calidad de las escuelas. Al empezar a declinar los puntajes en el decenio de los sesenta, los críticos los señalaron como un indicador de que la calidad de la educación estaba declinado. En el decenio de los ochenta, el Departamento de Educación de los E.E.U.U. comparaba anualmente los puntajes SAT de los estados para comparar su desempeño educacional.

Pero O'Reilly y otros investigadores dicen que tales comparaciones carecen de sentido. Dicen que esas comparaciones ignoran el hecho de que someterse al SAT es producto de una autoselección, y que el grupo que lo toma no es una muestra consistente que represente a la población como un todo. Dice O'Reilly: "Tratar de comparar escuelas exclusivamente sobre la base de sus puntajes en el SAT... pasa por alto muchas otras cosas que están sucediendo en las escuelas".

El Lanzamiento Federal

El SAT creó el escenario para el siguiente paso en el uso de pruebas. Mientras que el SAT, el C.I., y otras pruebas selectivas tuvieron altas implicancias para estudiantes individuales durante todo el siglo, hasta los años sesenta no habían evaluaciones que acarrearan consecuencias a la gente que conducía las escuelas.

Programas de pruebas como los de Iowa y Stanford habían predominado desde los años treinta, pero se pretendía que sus resultados se utilizaran para informar a los profesores sobre cómo instruir a los alumnos, no para evaluar cuán bien hacían su trabajo ellos mismos. "Solíamos tener pruebas de logros estandarizadas (en los años cincuenta), pero nunca fueron usadas para juzgar la calidad de la enseñanza", dice W. James Popham, un profesor de ciencias que se convirtió en un destacado experto en pruebas como catedrático en la Universidad de California en Los Angeles.

Antes de los setenta, las calificaciones de las pruebas se consideraban "sólo para uso interno" y rara vez eran comunicadas a funcionarios federales o estatales o al público, según escribe en un texto de 1989 sobre medición educacional Joy A Frechthling, entonces directora de responsabilización educativa de las escuelas del condado de Montgomery en Maryland.

Todo eso cambió cuando el gobierno federal empezó a desempeñar un papel cada vez más importante en la subvención de las escuelas y quiso ver los retornos a su inversión.

En 1965, la Oficina de Educación de los E.E.U.U. firmó un contrato con el sociólogo James S. Coleman para estudiar si las escuelas norteamericanas ofrecían las mismas oportunidades a los estudiantes blancos y negros.

El informe continúa siendo uno de los más grandes y significativos estudios sobre logros educacionales jamás realizados. Coleman y su equipo encuestaron a 570,000 estudiantes y 60.000 profesores de todo el país. Encontraron que los antecedentes familiares de los estudiantes y la estructura socio-económica de sus escuelas eran factores más significativos de los logros de los estudiantes que la calidad de sus escuelas.

Estos hallazgos han sido debatidos desde entonces, pero el estudio se convirtió en un prototipo para la realización de investigaciones educacionales que colocan las calificaciones en las pruebas en el centro del escenario. *"El informe Coleman redujo formalmente el asunto de cuán bien atendían las escuelas a estudiantes de las minorías o de bajos ingresos a un solo criterio, el desempeño del estudiante en pruebas objetivas de habilidades básicas, con preguntas de opción múltiple"*, escriben Baker y Stites.

Este supuesto empezó a orientar cómo el gobierno federal manejaba sus crecientes inversiones en educación básica. Después de que el presidente Lyndon B. Johnson firmara La Ley de Educación Primaria y Secundaria en 1965, su programa para ayudar a las escuelas con alta concentración de niños pobres pronto empezó a reflejar una definición de éxito ligada a las pruebas.

Para hacerse acreedores a fondos del programa *Title I*[2], dijo el gobierno federal, los distritos escolares tenían que mostrar resultados. El gobierno creó el Sistema de Evaluaciones e Información del *Title I* , también llamado TIERS (*Title I Evaluation and Reporting System*). Este exigía a las escuelas evaluar sus programas federales usando pruebas referidas a normas -- que comparan a los estudiantes con una muestra nacional -- y contribuyó a la "expansión sustancial" de su uso a lo largo de ese decenio, según un documento escrito en 1998 por Robert J. Linn. Él es co-director de CRESST (Centro de Investigación sobre Estándares y Evaluación de Estudiantes) de la Universidad de California en Los Ángeles y catedrático en Educación de la Universidad de Colorado en Boulder.

El TIERS animaba a las escuelas a evaluar a los estudiantes del *Title I* dos veces al año, escribe Linn, porque la mejor manera de demostrar el progreso académico era comparando el puntaje de un estudiante en el otoño con su puntaje en la primavera. Los estudios mostraban que las escuelas que seguían ese patrón mostraban mayor progreso de sus estudiantes que aquéllos que aplicaban las pruebas sólo una vez durante el año escolar.

[2] Título I (*Title I)* es un amplio programa del gobierno federal de los E.E.U.U. de apoyo a iniciativas de sus estados y localidades, de naturaleza compensatoria para grupos de escolares en riesgo, que incluye programas de instrucción remedial, de mejoramiento de la enseñanza, de introducción de innovaciones y otros – NT.

Nueve años después del inicio del *Title I*, el gobierno federal ordenó un tipo diferente de pruebas en una nueva ley sobre educación especial.

La Ley "Educación para Todos los Niños Minusválidos" de 1975 exigía a las escuelas examinar a los niños con problemas de aprendizaje para determinar si calificarían para servicios individualizados bajo ese programa. La ley incluso ordenaba que las escuelas evaluaran a todo niño potencialmente discapacitado dos veces, según escribe Linn en la edición de 1989 de **Medición Educativa**. Esa disposición fue incluida para asegurar precisión, pero, al igual que TIERS, duplicó el tiempo dedicado a las pruebas. También reforzó el rol de las pruebas de C.I. y otros instrumentos psicológicos diseñados para probar habilidad innata.

Competencia Mínima

Mientras tanto, entre los líderes federales de los años sesenta había ido creciendo el apoyo a una nueva prueba diseñada para proveer una imagen instantánea a nivel nacional del logro de los estudiantes.

En 1963, el Comisionado de Educación de los E.E.U.U., Francis Keppel, formó un panel de expertos para explorar formas de elaborar tal sistema de evaluación. En 1966, un año después de dejar su puesto federal, Keppel escribió: "La nación podía enterarse sobre edificios escolares o descubrir cuántos años permanecen los niños en la escuela, pero no tenía una manera satisfactoria de evaluar si el tiempo que pasaban en la escuela era efectivo".

Para presidir el comité, Keppel nombró a Ralph W. Tyler, quien había dirigido el Estudio de Ocho Años, evaluación sobre la educación secundaria progresiva, que marcó un hito histórico 30 años atrás.

El comité Tyler recomendó un muestreo regular de estudiantes en materias básicas para el programa, que vino a ser conocido como la Evaluación Nacional del Progreso Educacional *(National Assessment of Educational Progress – NAEP)*. Debido a que los administradores escolares locales objetaban que se reportaran resultados desagregados a nivel estatal, el comité Tyler propuso que los puntajes se reportaran para cada una de cuatro regiones. Según Maris A. Vinoskis, una historiadora de la Universidad de Michigan que escribió una historia del programa en 1998, si bien esta transigencia fue necesaria para calmar preocupaciones acerca de que la prueba federal pudiera convertirse en la base para las decisiones curriculares estatales y locales, también significó que los resultados de NAEP resultaran inútiles para evaluar la efectividad de las escuelas. La falta de datos estatales comparables obligó a los funcionarios federales a basarse en otros datos tales como los puntajes del SAT cuando, en los ochenta, elaboraron un afiche *(wall chart)* en el cual categorizaban a los estados. Como consecuencia, la tarea de evaluar a los distritos escolares recayó sobre los estados.

Empezando a inicios de los setenta, los estados se propusieron determinar lo mínimo que los estudiantes deberían saber antes de graduarse. Para evaluar si los individuos alcanzaban esos estándares básicos, los estados crearon las llamadas "pruebas de competencia mínima". Los niveles eran " lo más mínimo imaginable", según Popham, catedrático emérito de UCLA.

De 1973 a 1983, el número de estados con pruebas de competencia mínima creció de dos a 34, anota Linn en un ensayo sobre el crecimiento de la medición a lo largo de los últimos 50 años, publicado en 1998 por CRESST.

Una ley de Carolina del Norte de 1977, por ejemplo, instituía un sistema de pruebas que cumpliera con tres objetivos: asegurar que los graduados de secundaria "posean destrezas mínimas", que identifique sus fuerzas y debilidades y que haga a las escuelas responsables por lo que enseñan a sus estudiantes.

Al igual que las anteriores pruebas de C.I. y el SAT, las pruebas de competencia mínima suscitaron cuestiones sobre equidad racial. Anota Lynn que en 1977, el primer año de la prueba en Florida, 75 por ciento de los estudiantes blancos pasaron en el primer intento, comparado con 60 por ciento de los hispánicos y menos de 25 por ciento de sus contrapartes afro-americanas.

Según datos citados por Linn, veinte años más tarde, los estudiantes de minorías habían reducido la brecha pero no la habían cerrado. En 1984, 70 por ciento de los estudiantes negros las aprobaron en el primer intento, pero para entonces casi 90 por ciento de los blancos que rindieron la prueba la estaban aprobando. Desde entonces, la tasa de aprobación de los afro-americanos ha disminuido gradualmente, mientras que los estudiantes blancos han continuado obteniendo puntajes aproximadamente del mismo nivel que antes.

Una vez que los estados se comprometieron a adherir lo que hoy se llaman "altas implicancias" o "altos costos" *(high stakes)* a las evaluaciones, no retrocedieron. Si algo han hecho en los últimos dieciséis años, ha sido incrementar esos costos.

En 1983, un panel federal publicó el influyente informe **Una Nación en Riesgo**. Éste declaraba penosamente inadecuadas a las escuelas norteamericanas y convocaba a la búsqueda de modos de medir si los estudiantes habían dominado una currícula rigurosa.

La llamada despertó actividad, no sólo para añadir consecuencias significativas a las pruebas estatales, sino para aumentar su dificultad. La mayoría de los estados, tal como lo han hecho cada vez que el número de estudiantes que se someten a las pruebas aumenta dramáticamente, recurrió a pruebas estandarizadas, suscitando un debate sobre si las pruebas referidas a normas deberían ser usadas para decidir dar sanciones o recompensas a las escuelas o para decidir si los estudiantes serán promovidos o se graduarán.

"Una buena prueba referida a normas indicará detalladamente las fortalezas y debilidades de un niño, destreza por destreza", dice Maureen Di Marco, vice-presidenta de asuntos educacionales y gubernamentales de la compañía editorial Riverside, división de la Houghton Mifflin que distribuye las pruebas de Iowa. A nivel de conjunto, una prueba tal puede jugar un papel en decisiones de responsabilización "si es parte de un sistema, no su única característica", dice Di Marco, quien fuera importante consejera sobre educación del ex - gobernador de California, Pete Wilson. Añade: "Puede ser la característica predominante. Será la medición más fuerte, más objetiva".

En los últimos años del decenio de los ochenta y a inicios de los noventa, algunos estados experimentaron con "evaluaciones auténticas", basadas en portafolios de trabajos

de los estudiantes y en preguntas que les exigían escribir ensayos, aun en materias tales como ciencias y matemáticas, que tradicionalmente los evitaban.

California abandonó su programa después de que conservadores, Di Marco entre ellos, argumentaron que las preguntas de las pruebas se entrometían en las creencias personales de los alumnos, y tradicionalistas se quejaron de que un estudiante podía sacar buen puntaje en un problema de matemáticas escribiendo un ensayo de alta calidad pero sin llegar a comprender los principios matemáticos subyacentes. En 1996, después de notar que los puntajes de las evaluaciones basadas en portafolios no eran confiables a nivel de los individuos, Vermont añadió una prueba estandarizada para complementarlas. Kentucky también está añadiendo pruebas estandarizadas a su paquete de evaluación. Mientras tanto, Maryland continúa confiando en su sistema orientado hacia la ejecución o desempeño.

La dependencia en pruebas estandarizadas no ha cambiado mucho en los noventa, dicen los expertos, aunque los estados declaran que sus pruebas están alineadas con los [nuevos] estándares curriculares que ellos han adoptado. Sin embargo, tan generales son muchos de esos estándares, que los que se encargan de elaborar pruebas necesitan hacer apenas un poquito más que revisar productos ya en sus anaqueles para satisfacer las necesidades de los estados, dice Popham, quien ocasionalmente compite por contratos como consultor independiente en evaluación.

"El actual lema de 'alineamiento' es mayormente una farsa", dice Eva Baker, directora del Centro Federal de Evaluación de la UCLA. Dice Popham que "*frecuentemente, parecen versiones recalentadas*" de pruebas estandarizadas. *"La mentalidad que los editores conllevan cuando crean una prueba es lo que ellos saben, y ellos nada saben sobre instrucción"*. La prueba de Tejas, señala Popham, la prepara Harcourt Brace, una de las tres más importantes editoras de pruebas. La novena edición de la Prueba de Logros de Stanford de Harcourt Brace es usada en varios otros estados. La Oficina de Pruebas de California, ahora propiedad del Consorcio Mc Graw – Hill, está ahora a cargo del programa de pruebas de Kentucky.

Aun aquéllos que apoyan el concepto de alinear estándares de contenidos y pruebas dicen que el alineamiento no siempre es perfecto. *"Es todavía una pregunta abierta a qué se parece un sistema alineado"*, dice Matthew Gandal, director de estándares y evaluaciones de Achieve Inc., un grupo de líderes corporativos y estatales que presionan por mayores logros estudiantiles. "Retóricamente, todo el mundo está en eso. Pero, en la realidad, ¿cómo se puede distinguir?"

Pruebas de Altas Implicancias

Mientras los estados estaban elevando el nivel de dificultad de sus propias pruebas, el gobierno federal elevaba las implicancias del NAEP (Evaluación Nacional del Progreso Educativo). Empezando en 1990, la evaluación nacional empezó a producir puntajes para cada estado.

Para muchos estados, los resultados del NAEP conllevan un peso significativo, tanto en términos sustantivos como en materia de relaciones públicas. Los bajos resultados de la prueba de lectura del NAEP en California en 1994, por ejemplo, dio municiones a los críticos de la filosofía de aprendizaje integrado del lenguaje que había

abrazado el estado, que se vio forzado a desplazarse hacia la instrucción fonética. Este año, Kentucky ha tenido que defender sus mejorías, al parecer estadísticamente significativas, de críticos que dicen que las mejorías ocurrieron porque en 1998 el estado excluyó a un porcentaje más alto de alumnos que en 1994.

Si la propuesta del Presidente Clinton para nuevas pruebas nacionales voluntarias se convierte en una realidad, las pruebas NAEP podrían tener consecuencias individuales para niños del cuarto y octavo grados. El plan está siendo actualmente estudiado por el directorio del NAEP, pero es poco probable que supere opiniones en contra fuertemente enraizadas en el Congreso, tanto entre Republicanos como entre Demócratas. .

Ninguno de los acontecimientos recientes de Kentucky ni de California habrían sucedido a principios de siglo -- por el simple motivo de que los puntajes de las pruebas no se hacían públicos.

Por ejemplo, en 1925 Tejas emprendió lo que probablemente fue en esa época "una de las más extensas investigaciones sobre las operaciones de las escuelas y del Departamento de Educación de un estado jamás realizadas", según una historia sobre el rol de los estados en la educación compilada por el Consejo de Funcionarios Principales de las Escuelas de los estados.

El estudio de Tejas produjo un reporte sobre el acceso de alumnos y la estructura burocrática, pero no mencionó los puntajes que obtuvieron en las pruebas. Más aun, el libro, que data de 1969 y detalla la historia del rol del estado en la educación, no menciona el sistema de pruebas de ese entonces.

Pero hacia 1980 Tejas empezó a apoyarse en los puntajes para tomar decisiones importantes. Ese año, empezó a requerir pruebas de competencia mínima en lectura, matemáticas y escritura.

En 1990 el estado introdujo la Evaluación de Habilidades Académicas de Tejas (*Texas Assessment of Academic Skills - TAAS*). A diferencia de sus predecesores, los resultados del TAAS imponen consecuencias para todos en las escuelas: alumnos, profesores, administradores y miembros del Consejo. Sin embargo, los expertos se preguntan si programas como el TAAS deben ser el criterio principal para definir a las escuelas como exitosas.

Popham dice: "Tienen ítems que realizan un pésimo trabajo en lo que se refiere a medir la calidad de las escuelas. Lo que se está midiendo es aquéllo con que los chicos llegan a la escuela y no lo que allí aprenden". La mayoría de los expertos académicos advierten el peligro de fiarse de una única prueba para tomar decisiones críticas como el que un estudiante pase al siguiente grado o se gradúe de secundaria.

En un reporte de 1998, un comité de la Academia Nacional de Ciencias escribía: "El puntaje en una prueba, al igual que otras fuentes de información, no es exacta. Es un estimado de la comprensión o dominio del estudiante en un momento en particular. De allí se desprende que no se deben tomar decisiones educacionales de altas implicancias única o automáticamente sobre la base de los calificativos obtenidos en una sola prueba, sino tomando también en cuenta otra información relevante".

Los defensores de las pruebas están de acuerdo con que los puntajes de las pruebas no deben ser el único factor en las decisiones sobre responsabilización. Pero sí creen que pueden jugar un rol preponderante. Gandal dice: "Deben situarse al centro de las políticas de responsabilización. Ellas son uno de los únicos indicadores confiables sobre qué están aprendiendo los alumnos. Esto no quiere decir que no pueden complementarse muy bien con lo que maestros y las escuelas traen a la ecuación".

Pero es poco probable que estos llamados a la moderación influyan en los formuladores de políticas y en el público. Los formuladores de políticas están hambrientos de información que confirme que sus escuelas están teniendo éxito y se están apoyando en los resultados de las pruebas para una variedad de "libretas de notas" *(report cards)*, bonos para los profesores y penalidades para los funcionarios de escuelas cuyos estudiantes obtienen puntajes bajos.

El público -- desde los reporteros de los diarios hasta los agentes de bienes raíces -- usa información extraída de las pruebas como "*una herramienta maligna para sus propios intereses*", dice Sherman Dorn, un profesor asistente de historia de la educación en la Universidad de Florida del Sur.

Hoy en día, son los administradores de las escuelas quienes están en la mira. Dorn dice: "*La historia ha jugado una broma cruel a los administradores de escuelas. Ellos usaban esas pruebas para hacer un seguimiento a sus alumnos en formas algunas veces malignas… y ahora éstas se están utilizando contra ellos*".