Las pruebas de aprovechamiento estandarizadas
como instrumento de medición y político

María del Rosario Medina

RESUMEN
El uso de pruebas de aprovechamiento estandarizadas para que los sistemas escolares y las escuelas rindan cuentas por el aprovechamiento de sus estudiantes y determinar la calidad de la educación ha tomado mayor realce y poder con la implantación de ley *No Child Left Behind*. Los estatutos de esta ley aplican a los estados y territorios de los Estados Unidos de América, incluyendo a Puerto Rico. Este artículo presenta las características principales de las pruebas de aprovechamiento estandarizadas como un instrumento de medición y critica su uso como un instrumento político, según se perfila en esta ley. También reseña algunos de los resultados de las pruebas de aprovechamiento que se han administrado en Puerto Rico desde el año 2000. Pretende, en fin, provocar una reflexión acerca del uso inadecuado de los resultados de estas pruebas como indicador del aprovechamiento académico de los estudiantes y de la calidad del sistema escolar público.

Descriptores: Pruebas estandarizadas; Ley ***No Child Left behind;*** Pruebas de aprovechamiento; Aprovechamiento académico; Rendición de cuentas; Medición Educativa

ABSTRACT
The use of standardized achievement test results for accountability purposes, and to judge the quality of educational systems and schools, has been given importance and power through the No Child Left Behind Act. This law applies to the states and territories of the United States of America, including Puerto Rico. This paper discusses the main features of standardized achievement tests as measurement instruments, as well as critiques their use as a political tool under this law. A summary of the results of standardized achie-

vement tests administered in Puerto Rico since the year 2000, is also included. The intention is to promote reflection about the misuse of standardized achievement test results as indicators of student academic achievement and of the quality of the public school system.

Keywords: Standardized testing; Standardized tests; No Child Left Behind law; Achievement testing; Academic achievement; Accountability; Educational Measurement

El aprovechamiento estudiantil es el principal interés de los sistemas escolares. Las pruebas de aprovechamiento estandarizadas se han utilizado como instrumentos[1] para medirlo y determinar la calidad[2] de la educación pública en los Estados Unidos de América por muchos años. Varios documentos y leyes han impulsado esta práctica. En el año 1960, se estableció el *National Assessment of Educational Progress*[3] (NAEP) para medir el aprovechamiento del estudiantado e ilustrar tendencias en su ejecución a lo largo del tiempo. La ley *Elementary and Secondary Education* (ESA), aprobada en el año 1965 por el Congreso de este país, requería la evaluación de los programas educativos que recibían fondos bajo el Título I (i.e., programas destinados a estudiantes de familias con bajos ingresos económicos). Un sinnúmero de pruebas de aprovechamiento y aptitud se administraron como parte de esas evaluaciones. Luego, el gobierno estadounidense utilizó los resultados para ejercer mayor control sobre la adjudicación y renovación de los fondos a estos programas. De este modo, el NAEP y los requisitos de la ley ESA constituyen los primeros usos de las pruebas de aprovechamiento como instrumentos para monitorizar el rendimiento del estudiantado estadounidense (Koretz & Hamilton, 2006).

En los años ochenta, los resultados de las pruebas de aprovechamiento contribuyeron a propagar la percepción de que existía una crisis en la educación pública y había que reformarla. El bajo desempeño de los y las estudiantes estadounidenses en las pruebas estandarizadas nacionales (e.g., el *Scholastic Aptitute Test* y el NAEP), así como en las internacionales (e.g., *Pruebas Internacionales de Aprovechamiento Académico*), se utilizó como un símbolo de la precaria calidad de la educación. Se usaron las puntuaciones de estas pruebas como indicadores del éxito o fracaso de las escuelas públicas del país. Ante esta

situación, la National Commision on Excellence in Education (1983), por medio del documento *A nation at risk*, propuso la administración de pruebas de aprovechamiento como el vehículo principal para impulsar una reforma educativa. Además, planteaba las posibles consecuencias para los sistemas escolares según los resultados obtenidos. Así prospera la idea de que las pruebas son un medio para aumentar el aprovechamiento escolar. Ya al final de esta década, los y las estudiantes estadounidenses contestaban entre tres y nueve pruebas de aprovechamiento estandarizadas anualmente.

La siguiente década se caracterizó por el desarrollo de estándares. La aprobación de las Metas 2000 (*Goals 2000: Educate America Act of 1994*, http://www.ed.gov/legislation/GOALS2000/TheAct/intro.html) propulsó la llamada "reforma educativa sistémica" basada en dos tipos de estándares: (1) de contenido[4] o curriculares, los cuales expresan lo que el estudiantado debe aprender en las distintas asignaturas escolares y (2) de ejecución, que indican el desempeño esperado. Además, se contempló la posibilidad de construir y administrar una prueba de aprovechamiento a nivel nacional (Carnevale & Kimmel, 1997). La intención era promover cambios en el currículo, la enseñanza y la evaluación del aprendizaje y, por ende, mejorar el aprovechamiento estudiantil (Medina, 1992, 1998). Se asignaron fondos federales a los estados y sus agencias educativas para distintos proyectos con este fin. El proyecto *Puerto Rico Systemwide Systemic Intitiative (PR-SSI)*, administrado por el Centro de Recursos para la Ciencia e Ingeniería de la Universidad de Puerto Rico, fue uno de los que recibió el auspicio financiero de la *National Science Foundation* (Clune, 2001). Para esta época, también prolifera la idea de aplicar diversos instrumentos de *assessment* [5]para evaluar el aprovechamiento estudiantil, tanto en las salas de clases como en los sistemas educativos de los estados.

El *National Council of Teachers of Mathematics* (1989) fue la primera organización profesional que publicó estándares curriculares en los Estados Unidos de América. Luego, otras organizaciones, estados y territorios emularon este esfuerzo. Entidades académicas, como la *National Academy of Sciences* y la *National Academy of Education*, también recomendaron el uso de los estándares. En Puerto Rico, se desarrolló el primer conjunto de estándares curriculares para Matemáticas y Ciencia en el 1996 (Comisión para el Desarrollo de los Estándares Curriculares y de *Assessment* para la Ciencia y Matemática Escolar en Puerto Rico, 1996). Cuatro años después, el Departamento de Educación de Puerto Rico (DEPR) publicó una serie de nuevos están-

dares en todas las asignaturas escolares y éstos son los que siguen vigentes.

En la actualidad, los estados y territorios de los Estados Unidos de América cuentan con estándares de contenido en casi todas las materias escolares y con un sistema de rendición de cuentas (*accountability*) al público de la gestión educativa (información disponible en http://www.achieve.org). Estos sistemas dependen, primordialmente, de los resultados de las pruebas de aprovechamiento estandarizadas para demostrar que la inversión de los fondos públicos rinde el producto esperado (i.e., el aprovechamiento escolar). Conviene destacar que el movimiento de reforma educativa basado en los estándares promueve la uniformidad en el currículo y la enseñanza sin considerar la diversidad del estudiantado, los distintos contextos sociales y económicos de las escuelas y las condiciones de trabajo de la clase magisterial (Medina, 1992,1998). Además, presume que todas las maestras[6] conocen y utilizan los estándares en la planificación y práctica de la enseñanza de las distintas asignaturas.

Por otra parte, este movimiento demanda dos requisitos en las pruebas de aprovechamiento: (1) que estén "alineadas" con los estándares de contenido o curriculares (denominadas *standards-referenced test* como una modalidad de las pruebas con referencia a criterios, *criterion-referenced tests*, que se describen más adelante) y (2) que provean información acerca de la ejecución o las puntuaciones del estudiantado con referencia a estos estándares. A menudo, la descripción de los estándares o niveles de ejecución se traducen en las puntuaciones o los porcentajes de pase de las pruebas de aprovechamiento (en inglés, *cutoff score*). Estos requisitos adquieren un carácter compulsorio con la ley conocida como *No Child Left Behind* ([NCLB], Public Law 107-110, http://www.ed.gov/policy/elsec/leg/esea02/), la cual el Congreso de los Estados Unidos de América aprobó en el 2001. Esta ley enmienda y reautoriza la ley ESA.

Tanto las Metas 2000 como la ley NCLB proponen una reforma educativa basada en los estándares curriculares y de ejecución y sistemas de *assessment* en los estados y territorios estadounidenses. Sin embargo, la ley NCLB va más lejos ya que impone un sistema obligatorio y regulado. Además, está atado a una serie de castigos y recompensas para los sistemas escolares y las escuelas, a partir de los resultados que obtienen[7]. Aunque esta ley señala que las pruebas de aprovechamiento no son los únicos instrumentos, ni son suficientes, para medir los aprendizajes complejos y dispone que se usen múlti-

ples medidas, la mayoría de los estados han implantado un programa de pruebas de aprovechamiento estandarizadas (Public Law 107-110, Sec.1111(b)(3)(A); Sec. 1111(b)(3)(C)(i-iv),(vi)). Algunos estados, como Kentucky, Vermont y Maryland, optaron por la alternativa de un sistema de *assessment* de la ejecución (conocido como *performance assessment)*, que han abandonado, mientras que otros combinan varios instrumentos (Ferrara & DeMauro, 2006).

Este artículo destaca las características principales de las pruebas de aprovechamiento estandarizadas como un instrumento de medición y critica su uso como un instrumento político, según se perfila en la ley NCLB. En particular, se refiere a las pruebas que construye una compañía comercial contratada por la agencia educativa del estado. También se reseñan brevemente algunos de los resultados de las pruebas de aprovechamiento que se han administrado en Puerto Rico desde el 2000. Pretende, en fin, provocar una reflexión acerca del uso inadecuado de los resultados de estas pruebas como indicador del aprovechamiento académico del estudiantado y de la calidad del sistema escolar público.

La prueba de aprovechamiento estandarizada como instrumento de medición

Una prueba es una muestra representativa de ítems[8] relacionados con el *construct*, dominio o contenido que intenta medir. Una prueba de aprovechamiento o rendimiento (en inglés, *achievement test*) tiene como propósito medir parte del conocimiento adquirido por una persona que ha sido expuesta a un proceso educativo. En este caso, sirve para indicar qué el estudiantado logró aprender en cierta asignatura escolar por medio de sus respuestas a una muestra de ítems. Esta prueba puede incluir ítems para seleccionar o suministrar la respuesta y tareas de ejecución. Una o varias maestras, una compañía comercial, una entidad del estado o independiente pueden construir este tipo de prueba. No obstante, en este artículo me refiero a las que construye una compañía comercial contratada por la agencia educativa del estado y que sirven al propósito de informarle acerca del aprovechamiento del estudiantado que asiste a las escuelas bajo su jurisdicción.

Una prueba estandarizada (en inglés, *standardized test*) implica que se administra (o una forma equivalente) a una persona o grupo de personas siguiendo un procedimiento determinado. La estandarización se refiere a la uniformidad en las condiciones de administración. Con este fin, se prepara un manual o protocolo que incluye, como

mínimo, las instrucciones, los límites de tiempo, las formas de responder las posibles dudas o preguntas de las personas que contestan la prueba, la manera en que se va a repartir, recoger, organizar y manejar los materiales. De esta manera, se reducen las posibles discrepancias entre las personas encargadas de administrar la prueba en distintos lugares y momentos. Además, estas personas deben procurar que sus rasgos personales no influyan en la aplicación de la prueba (e.g., el tono y la inflexión de la voz, la expresión facial y su manera de vestir). Como he destacado en otra publicación, el ambiente físico o lugar donde se administra la prueba también debe tener unas condiciones apropiadas para evitar las interrupciones innecesarias y la intervención de otros factores, como el ruido (Medina-Díaz & Verdejo-Carrión, 2000). Cuando los y las estudiantes que pertenecen al programa de Educación Especial[9] contestan estas pruebas, las condiciones de administración se deben acomodar a sus necesidades particulares. Cuán bien se sigan los procedimientos de administración establecidos contribuye a la validez y confiabilidad de las puntuaciones de la prueba.

Aparte de unas condiciones de administración rigurosas, una prueba estandarizada conlleva, también, procedimientos uniformes de calificación e interpretación, que se establecen de antemano. La documentación técnica de la prueba debe indicar detalladamente cómo se van a calificar o valorar las respuestas, bien sea de forma manual o automatizada. En el caso de una prueba con ítems para seleccionar la respuesta (e.g., alternativas múltiples) se usa una clave que indica las respuestas correctas. Este procedimiento lo podría llevar a cabo una computadora, siempre y cuando tenga alguna manera de insumo de las respuestas y la programación adecuada para aplicar la clave y obtener los resultados. Si la prueba incluye preguntas para elaborar la respuesta u otras tareas de ejecución, entonces se necesita una rúbrica o matriz de valoración, un procedimiento específico para usarla y el personal debidamente adiestrado para aplicarla de manera confiable.

Por lo general, los resultados de una prueba de aprovechamiento estandarizada se interpretan haciendo referencia a un conjunto de normas o de criterios. Con bastante frecuencia, las pruebas de este tipo son normalizadas. Tal vez por esto se tiende a pensar que son sinónimos. Una prueba normalizada es la que ha pasado por un procedimiento de normalización; es decir, se han calculado las normas de la prueba a base de las respuestas de una muestra representativa de personas. Este grupo tiene características similares al que se destina la prueba y se le llama grupo normativo o de referencia. Por ejemplo, una prueba

de aprovechamiento de matemática para estudiantes de tercer grado se administra a una muestra grande de estudiantes de este grado en Puerto Rico con distintas características (e.g., género, área geográfica de residencia, raza, nivel socioeconómico) para determinar la distribución de sus puntuaciones. Con estos datos, se calculan las normas o estadísticas (descriptivas y derivadas) que indican la ejecución típica o "normal" de la muestra (o grupos) de personas que contestó la prueba (e.g., la media aritmética, el rango porcentual y la equivalencia en grados). Para interpretar la puntuación o ejecución de un estudiante en la prueba, se compara ésta con las estadísticas que se calcularon del grupo de referencia. En algunas ocasiones, hay más de un grupo normativo o en el mismo grupo se identifican diferentes variables para hacer ciertas comparaciones (e.g., género, edad, nivel escolar, área geográfica de residencia). Además, si los ítems son de alternativas múltiples, se seleccionan aquellos que tienen mayor poder de discriminación entre los y las estudiantes (i.e., aquellos que pueden contestar correctamente cerca de la mitad del estudiantado). Por el cuidado en la selección del contenido y de los ítems es que estas pruebas pueden discriminar entre el estudiantado con alto y bajo aprovechamiento, y se pueden comparar sus puntuaciones (Medina-Díaz & Verdejo-Carrión, 2000).

En una prueba cuyas puntuaciones se interpretan con referencia a unos criterios establecidos, hay una cantidad adecuada y suficiente de ítems (por lo menos cinco) que representan los estándares curriculares (pueden ser también objetivos instruccionales o destrezas) y se determina la cantidad de ítems correctos o la ejecución que se considera aceptable para demostrar dominio. A ésta se le llama puntuación de pase o nivel de ejecución. Cuando se usan los estándares curriculares como base para la construcción de la prueba, se espera que en la interpretación de las puntuaciones (individuales y grupales) se haga referencia a los mismos. Por ejemplo, se debe indicar la cantidad de ítems correctos o la puntuación que el estudiantado obtuvo en cada estándar.

Desde la década pasada, las agencias educativas (e.g., DEPR) de los estados y territorios de los Estados Unidos de América han contratado compañías comerciales[10] para construir pruebas de aprovechamiento con distintos propósitos. Se espera que los ítems de estas pruebas representen los estándares curriculares particulares que los estados y territorios desarrollaron. Por esta razón, las compañías tienen que construirlas "a la medida" (en inglés, *custom-developed test* o *tailored tests*) de las exigencias de la agencia educativa que las contrata. Lo que

implica, por supuesto, un costo más alto en este tipo de prueba que en otra que la compañía ya haya desarrollado. Este requisito ya levanta interrogantes acerca de la facilidad y responsabilidad de las compañías para producirlas. Además, provoca inquietudes acerca de la capacidad y los recursos de las propias agencias educativas de los estados y territorios para monitorizar y evaluar el servicio de las compañías, así como la calidad técnica de las pruebas.

Conviene destacar que las maestras y las directoras escolares no son responsables de la selección de las pruebas de aprovechamiento estandarizadas, sino la agencia educativa. La participación del magisterio en el diseño y la construcción de estas pruebas son mínimas. No obstante, es posible que las compañías consulten y empleen a algunas maestras en las distintas etapas de la construcción. Por lo general, el proceso de construcción de estas pruebas incluye los siguientes pasos: (1) seleccionar los estándares que se van a incluir en la prueba; (2) preparar las especificaciones; (3) escribir las ítems que representen cada estándar; (4) revisar y editar los ítems; (5) preparar los procedimientos o medios para la calificación de las respuestas; (6) establecer los procedimientos y redactar las instrucciones para la administración; (7) administrar los ítems de manera "piloto" a una muestra representativa de la población que contestará finalmente la prueba; (8) analizar las respuestas; (9) seleccionar y revisar las ítems para la versión final de la prueba; (10) administrar la prueba a una muestra representativa de estudiantes del estado; (11) desarrollar normas o estándares de ejecución; (12) recopilar evidencia de la validez de las puntuaciones e inferencias; (13) preparar los informes de los resultados y (14) escribir la documentación técnica.

En fin, una prueba estandarizada bien construida conlleva tiempo, peritaje y gastos. Se espera que la compañía que realiza este trabajo cuente con los recursos humanos, tecnológicos y fiscales que garanticen el cumplimiento de los estándares de calidad para las pruebas educativas y psicológicas (i.e., *Standards for Educational and Psychological Testing*) publicados por *American Educational Research Association, American Psychological Association y National Council of Measurement in Education* (AERA, APA & NCME,1999). Se confía, además, que quienes construyen la prueba documenten, de manera detallada, cómo se diseñó y provean evidencia de la validez y confiabilidad de las puntuaciones. Esta información, por lo general, aparece en el manual técnico de la prueba.

La ley No Child Left Behind

La ambiciosa ley NCLB incluye cambios substanciales en el control del gobierno de los Estados Unidos de América sobre los sistemas educativos de sus estados y territorios. Las secciones principales del Título I, Parte A de la ley NCLB (Public Law 107-110, Title I) requieren que cada estado mediante su agencia educativa[11]:

a) adopte estándares de contenido y de desempeño académico (*challenging academic content standards and challenging student achievement standards*) que apliquen a todas las escuelas y el estudiantado (Sec.1111(b)(1)(A) y Sec.1111(b)(1)(B));
b) desarrolle e implemente un sistema de *accountability* que sea efectivo en asegurar que todas las agencias educativas y escuelas públicas elementales y secundarias alcancen el progreso anual adecuado (*academic yearly progress* [AYP]). Este sistema se debe basar en los estándares y *assessments* académicos y otros indicadores académicos; debe tomar en consideración el aprovechamiento de todo el estudiantado en las escuelas públicas elementales y secundarias e incluir sanciones y recompensas, tales como bonos y reconocimientos para asegurar que las agencias educativas y escuelas se responsabilicen del aprovechamiento estudiantil y se aseguren de cumplir con el AYP (Sec.1111(b)(2)(A)(i) y (iii));
c) implemente un conjunto de instrumentos de *assessment* académico del estudiante cada año y de alta calidad (*a set of high-quality, yearly student academic assessments* en el texto de la ley); que incluya como mínimo, Matemáticas, Lectura o Artes del lenguaje y Ciencia. Estos se usarán como medios para determinar el desempeño anual del estado, de cada agencia educativa local y escuela en capacitar a los estudiantes para el logro de los estándares. Para el año 2005-2006, se debe medir el aprovechamiento, relativo a los estándares de contenido y de ejecución, de todos los estudiantes de tercer a octavo grado y un grado de escuela superior en lectura o artes del lenguaje en Inglés y Matemáticas (Sec.1111(b)(3)(A)(vii)).
d) provea los acomodos razonables y adaptaciones para el estudiantado con necesidades especiales y se incluyan estu-

diantes con habilidad limitada en Inglés[12] (Sec.1111(b)(3)(A)(ix)(II) y (III)).

e) mida la proficiencia[13] (*proficiency*) del estudiantado (de tercer a quinto, de sexto a noveno, y de décimo a duodécimo grado) en Ciencia, comenzando en el año 2007-2008 (Sec.1 111(b)(3)(A)(v)(II)).
f) demuestre qué constituye el AYP[14] hacia el desempeño esperado en estas materias (Sec.1111(b)(3)(C)).
g) defina un AYP que: aplique los mismos estándares de ejecución para todo el estudiantado; mida el progreso de las escuelas elementales y secundarias y de la agencia educativa local; incluya objetivos anuales, por separado, para el progreso continuo y sustancial del aprovechamiento de todo el estudiantado y de otros grupos (estudiantes con escasos recursos económicos, de grupos raciales y étnicos; con necesidades especiales y con habilidad limitada en Inglés); incluya la tasa de graduación de escuela superior y otros indicadores académico que considere pertinentes (Sec.1111(b)(3)(B)y (C)(i- vi));
h) identifique las escuelas que no alcanzan el AYP y los estándares mínimos de desempeño, denominadas en la ley como en necesidad de mejoramiento (*in need of improvement*, *school improvement*) y les provea asistencia técnica para implementar el plan de mejoramiento, acción correctiva o reestructuración (Sec. 1116 (a)(1)(B), (4), (7) y (8)).
i) divulgue anualmente, a las madres, los padres, las maestras, las escuelas y la comunidad, los resultados de las pruebas u otras técnicas de *assessment* así como del progreso del estado y las escuelas en cumplir con el AYP (Sec. 1116 (a)(1)(A) y (C)).
j) desarrolle un plan que asegure que, al finalizar el año 2005-2006, todo el magisterio y las maestras en las asignaturas académicas medulares (*core academic subject*) estén altamente cualificados (*highly qualified*)(Sec.1119(a)(2)); y
k) prepare y divulgue un informe anual (*annual state report card*) conciso y en un formato uniforme y fácil de entender (Sec.1111 (9)(h)(1)(A)).

Para al año escolar 2013-2014, se espera que *todo* (100%) el estudiantado alcance o sobrepase el nivel de proficiencia en las pruebas de aprovechamiento estandarizadas de Lectura o Artes del Lenguaje

y Matemáticas que se administran en los estados y territorios. Estos tienen que demostrar cada año su progreso en alcanzar el nivel de aprovechamiento esperado y así seguir recibiendo los fondos federales. Concurro con Linn (2005) en que esta expectativa refleja una aspiración política más que una aspiración real ya que es contraria a lo que el sentido común y la investigación han demostrado. A mi juicio, es imposible lograrla por las siguientes razones: (a) las puntuaciones bajas en las pruebas de los distintos grupos de estudiantes, (b) las dificultades de los estados y territorios en la implementación de lo que dispone la ley, (c) la variedad de medidas y métodos que se utilizan para calcular el AYP y (d) el aumento leve en las puntaciones del NAEP.

Abedi y Dietel (2004) advierten que alcanzar dicha meta es casi imposible para los y las estudiantes cuyo lenguaje materno no es inglés[15] y quienes lo están aprendiendo en la escuela (llamados *English Language Learners*). Estos estudiantes, así como los afroamericanos, latinos y de bajos recursos económicos han tenido, de manera consistente, puntuaciones bajas en las pruebas de aprovechamiento estandarizadas que se han administrado en los estados. Parece que en vez de reducirse, continúa ampliándose la diferencia en el aprovechamiento (en inglés, *achievement gap*) entre los y las estudiantes de estos grupos y sus contrapartes anglosajones y de mayores recursos económicos. Debido a la presión que han ejercido los administradores, los educadores, los padres, las madres y las organizaciones profesionales, el Departamento de Educación de los Estados Unidos comenzó, en el 2004, a ser más flexible con la participación e interpretación de las puntuaciones de los y las estudiantes en el programa de Educación Especial y cuyo lenguaje materno no es inglés.

De igual forma, varios artículos periodísticos, así como informes gubernamentales y de investigación indican que los estados han tenido grandes dificultades en la implantación de las disposiciones de la ley. Linn (2003), Hamilton y Stecher (2004) y Tucker y Toch (2004) destacan que la mayoría de los estados no están preparados para atender lo que esta ley exige, así como para ofrecer el tipo de apoyo y recursos humanos y financieros que requieren las escuelas para llegar al AYP. Recientemente, el estudio de Vázquez Pérez (2006) y la prensa de Puerto Rico (Roldán Soto, 29 de enero de 2006) también señalan la negligencia del DEPR en proveer la "ayuda" que requiere la ley NCLB para algunas escuelas en "plan de mejoramiento". En los años 2003 y 2004, varios estados (Kansas, Kentucky, Luisiana, Missouri, Montana, Nebraska, Carolina del Sur y Wisconsin) sometieron demandas legales

contra el gobierno federal reclamando fondos adicionales para que las escuelas en las comunidades de escasos recursos económicos puedan recibir la instrucción apropiada que se requiere para contestar bien las pruebas (Dobbs, 2004).

Más aún, hay diferencias notables entre los estados y territorios en los sistemas de *accountability*, las pruebas y otros instrumentos de *assessment* que utilizan, así como en los grados y las asignaturas en que se administran. Los estados y territorios también han establecido sus propias metas, niveles de ejecución y métodos de informar los resultados y calcular el AYP (Chester, 2005). Además, es posible que las agencias educativas recurran cada año a compañías externas para realizar los cómputos del AYP, como es el caso de Puerto Rico (entrevista con Ángel Canales, 24 de enero de 2007). Todos estos factores contribuyen a explicar la variabilidad en la cantidad de escuelas que logran o no el AYP (Ferrara & DeMauro, 2006; Linn, Baker & Herman, 2005; Choi, Goldschmidt & Yamashiro, 2005). Golberg (2005) señala, que en el 2004, aproximadamente 26,000 (28%) de las 91,400 de escuelas públicas en los Estados Unidos de América no lograron el AYP establecido. Para el mismo año, el Departamento de Educación aprobó totalmente los planes de mejoramiento del aprovechamiento escolar de 28 estados y condicionalmente los de 23. La aprobación de estos planes revela la intensión de la ley NCLB de ejercer control y poder sobre las operaciones de las agencias educativas de los estados y territorios.

Además, la ley requiere que las escuelas y las agencias educativas reporten por separado los resultados en las pruebas de los estudiantes de Educación Especial, con desventajas económicas, cuyo lenguaje materno no es inglés y por raza y origen étnico e impone penalidades si estos subgrupos de estudiantes no demuestran progreso. Los estados y territorios pueden decidir la cantidad y los subgrupos de estudiantes de los que presentarán los resultados. El porcentaje de estudiantes proficientes en una escuela (i.e., los que obtienen puntuaciones en o sobre una puntuación establecida que indica el nivel de proficiencia) es la estadística que comúnmente se informa. Se emplea la diferencia en el porcentaje de estudiantes que logran el nivel de proficiencia cada año como un indicador de ganancia. Sin embargo, este porcentaje es una estadística que indica tendencia de las puntuaciones del estudiantado y no refleja cambios o crecimiento a lo largo del tiempo, ya que no se consideran los mismos estudiantes cada año. Koretz y Hamilton (2006) destacan los errores en los procedimientos para calcular estas diferencias y posibles maneras de remediarlos. También, cabe la posi-

bilidad de que cada año aumente el porcentaje de escuelas que logran los AYP debido a la manipulación de los datos y la eliminación de grupos de estudiantes. Goldschmidt y Choi (2007) ilustran los beneficios de un cómputo basado en el mejoramiento del estudiantado y de las escuelas de manera longitudinal.

Para evitar la corrupción y manipulación de los datos, la ley NCLB dispone que los resultados de las pruebas de aprovechamiento se comparen con los del NAEP. Desde el año escolar 2002-2003, una muestra de estudiantes de cuarto y octavo grado contestan las pruebas de lectura y Matemáticas del NAEP. De esta manera, los estados y territorios proveen evidencia adicional acerca de la precisión y certeza de sus informes del AYP. Esta prueba se administra cada dos años. En el 2003, ningún estado o distrito escolar había logrado que 100% del estudiantado de cuarto y octavo grado alcanzaran el nivel básico en la pruebas de lectura y de Matemáticas del NAEP (Linn, 2005). Linn, Baker y Herman (2005) encontraron disparidades entre los porcentajes de escuelas que alcanzaron el AYP en 46 estados y los porcentajes de estudiantes proficientes en lectura que reportaron en el siguiente año. Los resultados de la prueba de lectura que se administró en el 2005 indicaron un aumento y una reducción promedio de sólo un punto (en una escala de 0 a 500) en las puntuaciones de los y las estudiantes de cuarto y octavo grado, respectivamente, en comparación con los y las que contestaron la prueba en el 2003.

El análisis de los resultados de los 50 estados, el Distrito de Colombia y las escuelas del Departamento de Defensa indica que las puntuaciones del estudiantado de octavo grado en el NAEP no aumentaron. Sin embargo, hubo un leve aumento en el porcentaje de estudiantes de cuarto y octavo grado clasificados como proficientes en el 2005. Los resultados de las pruebas de Matemática son similares. Las puntuaciones promedios del estudiantado en cuarto y octavo grado aumentaron en tres y un punto, respectivamente, en comparación con las del 2003. Las puntuaciones promedios de los grupos de estudiantes afroamericanos y latinos en ambas pruebas aumentaron levemente en la administración del 2005. (Los informes de los resultados están disponibles en la página electrónica http://nces.ed.gov/nationsreportcard/).

En estos informes no aparece información relacionada con los resultados en Puerto Rico. Sólo se encontraron datos del país en la sección de *State profiles*. Estos indican algunas de las características de nuestras escuelas públicas en los años 2003-2004 y 2004-2005. Sin

embargo, el NAEP se administró por primera vez en Puerto Rico en el 2003. La prueba de Matemática se administró a una muestra al azar de 3,000 estudiantes de cuarto y octavo grado de 100 escuelas públicas (Rivera Sánchez, 30 de marzo de 2007). Sólo 9% y 4% de los estudiantes de cuarto y octavo grado, respectivamente, alcanzaron un nivel básico de los conocimientos y destrezas de Matemática. Los resultados de la muestra de estudiantes que contestaron la prueba en el 2005 revelan un leve aumento: 12% y 6% de los estudiantes cuarto y octavo grado, respectivamente, lograron el nivel básico (Rivera Sánchez, 30 de marzo de 2007).

La prueba de aprovechamiento estandarizada como instrumento político

Las pruebas de aprovechamiento estandarizadas siempre han sido objeto de múltiples críticas, como: (a) la limitada o poca representatividad del contenido curricular, de los estándares, de los aprendizajes cognitivos, psicomotores y afectivos importantes; (b) las limitaciones técnicas (e.g., la pobre evidencia de la validez de las puntuaciones y los sesgos entre los grupos con distintos trasfondos socioculturales y económicos); (c) los efectos en el estudiantado, las maestras y los directores escolares (e.g., la ansiedad, la frustración, la vergüenza y la deshonestidad académica) y (d) el mal uso de los resultados (e.g., la clasificación y la competencia de las escuelas). A partir de la aprobación de la ley NCLB, los ataques han sido más agudos por las consecuencias severas que tienen los resultados en las escuelas, el magisterio y el estudiantado (Wallis & Steptoe, 2007; Meier & Wood, 2004; Kohn, 2004; Sadker & Zittleman, 2004; Neill, 2003 a,b; Hughes & Bailey, 2001/2002; Nathan, 2002). A continuación destaco algunos de los asuntos relevantes de estas críticas y que merecen considerarse cuando se adjudican a estas pruebas el "poder" que no tienen.

Representatividad limitada del contenido curricular

Estas pruebas intentan cubrir el contenido y los estándares curriculares de la asignatura y el grado del sistema escolar de un estado o territorio. Por lo tanto, el énfasis no es en el contenido específico que ha enseñado una maestra de un grado o asignatura ni de un libro de texto o material curricular particular. Proveen una visión amplia y externa de lo que el estudiantado ha aprendido en una asignatura escolar y en un momento determinado. Tampoco se consideran las prácticas de enseñanza que se utilizan en las distintas escuelas, ni las diferencias

en los contextos socioeconómicos, las oportunidades educativas y las características de las maestras y de los grupos de estudiantes.

Si la prueba es normalizada, la muestra de ítems que incluye de cada tema de contenido o estándar curricular es pequeña, lo que no permite identificar, de manera certera, las fortalezas y dificultades de cada estudiante. Tampoco es probable que ofrezca los detalles necesarios para planificar la enseñanza en la sala de clases. Para esto se utilizan las pruebas y otros instrumentos de *assessment* que las maestras construyen y administran a sus estudiantes. Por el contrario, si las puntuaciones se interpretan a base de criterios (i.e., "el dominio de los estándares") se debe indicar la cantidad o el porcentaje de ítems correctos en cada uno de los estándares que cubre la prueba, de manera individual y grupal (por grado, escuela, distrito, región y estado). Para esto, es necesario que la prueba cuente con una cantidad de ítems adecuada para representar cada estándar curricular. Además, requiere comparar las puntuaciones individuales y grupales con los estándares de ejecución establecidos. Debido a que este tipo de interpretación provee mayor información del desempeño de los y las estudiantes, se espera que los resultados de la prueba ayuden a mejorar su rendimiento. Sin embargo, la manera en que se informan los resultados y en la fecha en que llegan a las escuelas y a manos de las maestras (al principio del próximo año escolar) impiden que puedan hacer cambios o ajustes en la enseñanza para los y las estudiantes que contestaron la prueba en el grado anterior.

En ambas interpretaciones, la puntuación de cada estudiante en la prueba se debe considerar como una aproximación de su conocimiento del contenido o "dominio" del estándar representado. La puntuación indica la cantidad de respuestas correctas (si es una prueba con ítems objetivos) o la calificación que se adjudicó (si es una prueba con tareas de ejecución) pero no dice porqué se obtuvo dicha puntuación. Por lo tanto, no se les puede adscribir a estas pruebas mayor "poder" del que tienen. Sólo provee una visión parcial, limitada y en un momento determinado de lo que el estudiantado conoce del contenido incluido. No hay prueba o instrumento de medición que sea capaz de medir o capturar la amplitud y complejidad de lo que se ha aprendido.

Tampoco se les puede adjudicar la facultad de evaluar la calidad de la educación de un estado, territorio o país. Es demasiado pretender que la ejecución o la puntuación en una prueba significa "calidad educativa". Como mencioné antes, una prueba es un instrumento que provee información limitada acerca del aprovechamiento estudiantil

en ciertas asignaturas escolares. Pophan (1996) también subraya que los resultados de estas pruebas no son indicadores apropiados de la calidad educativa ya que: (a) no proveen un indicador válido de la efectividad instruccional, (b) no hay correspondencia entre lo que se enseña en las escuelas y lo que se incluye en las pruebas, e (c) incluyen ítems que reflejan posibles sesgos en los aprendizajes esperados y las experiencias educativas a los que están expuestos los distintos grupos de estudiantes. Respecto a este último aspecto, se debe presentar evidencia empírica de la certeza y la precisión de la prueba para proveer información acerca de la ejecución de los diversos grupos de estudiantes en los distintos contextos escolares.

Limitaciones técnicas

Aún cuando a estas pruebas se les atribuyen propiedades que las convierten en instrumentos que cumplen con los requisitos de calidad técnica, deben demostrarlo. Tanto las compañías comerciales que desarrollan estas pruebas como las agencias educativas y otros usuarios son responsables de que haya evidencia para asegurar que las pruebas miden lo que reclaman (AERA, APA & NCME, 1999; Joint Committee on Fair Testing, 2004). La calidad de un instrumento de medición se demuestra de manera teórica, lógica y empírica con información que documenta la validez, la confiabilidad y el uso apropiado de las puntuaciones e inferencias (AERA, APA & NCME, 1999). Por ejemplo, la alineación de estas pruebas y otros instrumentos de *assessment* con los estándares curriculares es un asunto crítico para evidenciar la validez de contenido. Si las pruebas no están alineadas apropiadamente a los estándares curriculares, no se pueden considerar como medidas válidas ni las inferencias que se hacen acerca del "dominio de los estándares".

Sin embargo, la ansiada "alineación" de las pruebas con los estándares curriculares es otra parte de la retórica política en vez de una realidad. No se reconocen las ambigüedades y los posibles problemas conceptuales y técnicos que acarrea. Hay diversas definiciones de lo que constituye alineación así como variedad de métodos para determinarla (e.g., Webb, 2006; Resnick, Rothman, Slattery & Vranek, 2003-2004; Bhola, Impara & Buckdenhal; 2003; Porter, 2002). Así que el método que utiliza la compañía o persona que desarrolla la prueba refleja su concepto de alineación. Más aún, no es posible obtener una correspondencia perfecta entre la prueba y los estándares curriculares ya que, para construirla, se requiere una muestra amplia y calibrada de

ítems. Tampoco existe un criterio que indique un nivel de alineación aceptable (Koretz & Hamilton, 2006). La claridad y especificidad de los estándares curriculares son otras variables que se deben considerar en el proceso de alineación.

Para ilustrar la alineación, por lo general, las compañías que desarrollan las pruebas presentan una tabla de especificaciones que indica que la mayoría de los estándares están representados en un ítem, como mínimo, y que la mayoría de los ítems parean con al menos un estándar. Sin embargo, esta información no es suficiente para demostrar la alineación de los estándares curriculares con los ítems por separado y con toda la prueba. La alineación requiere un proceso en dos direcciones o vías: de los ítems a los estándares y viceversa, para identificar las posibles omisiones. Además, conlleva un proceso complejo de análisis y juicio de expertos[16], tanto de los ítems como la prueba, en las distintas dimensiones del contenido y la demanda cognitiva para contestar. Así que se requieren estudios y métodos más profundos para evaluar la alineación.

A modo de ejemplo, Resnick, Rothman, Slattery y Vranek (2003-2004) aplicaron un protocolo con varios criterios (centralidad del contenido, centralidad de la ejecución, reto y balance/amplitud) para evaluar la alineación entre los estándares curriculares, los ítems y las pruebas de aprovechamiento que se administran en cinco estados. Encontraron que los ítems, individualmente, están bien alineados con los estándares, pero las pruebas no logran medir la gama de estándares curriculares que los estados tienen. Los estándares que incluyen el contenido y los procesos cognitivos de mayor reto están representados de manera limitada o se omiten en las pruebas. Hay una mayor cantidad de ítems que conllevan procesos cognitivos simples para contestarlos. Los hallazgos de este estudio subrayan lo que he mencionado anteriormente: las compañías que desarrollan las pruebas, así como los funcionarios de las agencias educativas y otros usuarios no pueden ignorar las limitaciones de las pruebas ni pretender que representan los estándares curriculares sin evidencia que lo demuestre. Una inspección visual para "determinar" la correspondencia de los ítems con los estándares no constituye evidencia suficiente para reclamar dicha alineación.

Dos asuntos técnicos relacionados con la validez de las inferencias de la ejecución de los estudiantes también merecen atención: (1) la precisión de la escala que se utiliza para establecer el nivel de proficiencia y clasificar el estudiantado y (2) la participación y el acomodo

del estudiantado con necesidades especiales. La precisión de la escala requiere examinar la certeza en que los niveles de proficiencia (proficiente, avanzado y básico) establecidos por la ley representan, con el mínimo error posible, la ejecución individual y grupal de quienes contestaron la prueba. Además, se debe analizar la consistencia en que el estudiantado se clasifica de la misma manera o exhiben una ejecución similar en otros instrumentos de *assessment* basados en el mismo contenido. Hasta el momento, la evidencia de estas pruebas para lograr esto es pobre. Hoover (2003), por ejemplo, encontró que hay gran variabilidad en las clasificaciones del estudiantado como proficiente o avanzado entre las pruebas de aprovechamiento y el NAEP. Una de las pruebas clasificó 13% de los estudiantes de tercer grado como proficientes o avanzados mientras que otra clasificó el 56%. Las inconsistencias pueden ocurrir por las formas en que describen los niveles de ejecución, los métodos para establecer los estándares de ejecución o la puntuación de pase y la composición del panel de jueces que las compañías utilizan.

La meta de incrementar la participación del estudiantado del programa de Educación Especial en la administración de las pruebas se ampara en la idea de aumentar la validez de las puntuaciones e inferencias a partir de los grupos agregados. Sin embargo, en la práctica esta idea tiene varios tropiezos. Primero, la falta de consistencia en la identificación y clasificación de estos estudiantes. Por ejemplo, hay discrepancias entre la cantidad de estudiantes que reportan las escuelas y la que tiene el DEPR (Entrevista Ángel Canales, 24 de enero de 2007). Al parecer, las administradoras escolares "escogen" los estudiantes de Educación Especial que van a contestar las pruebas. Segundo, no hay documentación disponible del tipo de acomodo que reciben en la administración de las pruebas y si corresponde a lo que establece cada Plan de Enseñanza Individualizada. Tercero, la cantidad y la heterogeneidad de los tipos de necesidades especiales de estos estudiantes demanda acomodos particulares, que a las escuelas y agencias educativas se les hace difícil de cumplir (e.g., administraciones separadas con más tiempo). Estos factores también afectan la validez y credibilidad de las puntuaciones y el AYP de las escuelas.

Desafortunadamente, algunos sistemas escolares carecen de los recursos fiscales y el peritaje técnico en el campo de la Medición Educativa para solicitar o llevar a cabo los estudios necesarios para verificar la alineación con los estándares curriculares que proclaman, así como de otros factores que pueden afectar la validez y confiabilidad

de las puntuaciones de las pruebas. En un estudio reciente, la Oficina de Contraloría del Gobierno de los Estados Unidos (Goverment Accounting Office [GAO]) encontró que 25 de los 38 estados que participaron no tenían evidencia adecuada de la validez y confiabilidad de las puntuaciones de las pruebas de todos los estudiantes (GAO, 2006). El DEPR también comparte esta deficiencia.

Efectos de las pruebas

Distintas fuentes destacan los efectos de las pruebas de aprovechamiento estandarizadas. Organizaciones profesionales, como la American Evaluation Association, la American Research Association y National Center for Fair and Open Testing han publicado declaraciones en contra del uso inadecuado de estas pruebas para tomar decisiones que afectan al estudiantado, el personal escolar y los sistemas escolares. Los estudios de Smith y Rottenberg (1991), Jones, Jones, Hardin, Chapman, Yarbrough y Davis (1999), Hughes y Bailey (2001/2002) documentan los siguientes efectos: (1) la reducción en el tiempo instruccional para dedicarlo a la "preparación para las pruebas"; (2) el énfasis en la enseñanza del contenido que incluyen las pruebas de Matemáticas y Lectura; (3) el menosprecio a las otras asignaturas escolares; (4) el uso de estrategias instruccionales que se asemejan a los de las pruebas; (5) la organización del estudiantado en grupos homogéneos; (6) la conversión de las maestras en autómatas; (7) la presión, tensión y ansiedad en las maestras; (8) la ansiedad en las pruebas y falta de confianza en las puntuaciones por parte del estudiantado; y (9) la compra de materiales relacionado con las pruebas. Otros autores también han denunciado que los grupos de estudiantes afroamericanos, latinos, asiáticos, de escasos recursos económicos y con necesidades especiales son los más afectados con los resultados de estas pruebas (Berliner, 2006; Darling-Hammond, 2004; Nathan, 2002; Townsend, 2002; Neill, 2003 a,b; Herman & Golan, 1993).

El uso de los resultados de estas pruebas como indicador del conocimiento logrado por los estudiantes también ha desenmascarado otra cruda realidad: que en algunas escuelas se enseña para aprobar las pruebas, en vez de fomentar el aprendizaje del contenido curricular relevante del grado o la asignatura. Ante la presión para aumentar las puntuaciones, el personal docente altera y reduce el currículo para concentrarlo en el contenido, las destrezas y los ítems que se incluyen en las pruebas (i.e., dirigiendo la enseñanza hacia la prueba, en inglés *teaching to the test*). Al parecer estas prácticas son más frecuentes en

las escuelas elementales por los efectos que tienen en el magisterio y el estudiantado (Wallis & Steptoe, 2007; Pedulla, Abrams, Madaus, Rusell, Ramos & Miao, 2003; Herman & Golan, 1993; Smith & Rottenberg, 1991). Además, ha conducido a comportamientos deshonestos administradores escolares, maestras, maestros y estudiantes para aumentar las puntuaciones (Haladyna, Nolen & Hass, 1991; Cizek, 2001; Harrigton-Lueker, 2000). Aunque en Puerto Rico no contamos con mucha información acerca de estos efectos en las escuelas, las reseñas noticiosas de Rosario (10 de mayo de 2004) y Caquías Cruz (7 de mayo de 2005) sugieren que las puntuaciones bajas en las pruebas podrían incitar a la corrupción y deshonestidad del personal escolar.

Mal uso de los resultados

Las críticas anteriores atacan, primordialmente, los requisitos técnicos de las pruebas. Sin duda, se necesita mayor cúmulo de investigación para corregir las dificultades que se plantean. El más severo, a mi juicio, y difícil de erradicar es el "mal uso de los resultados de las pruebas" y la confianza excesiva en las pruebas. Esto se manifiesta de maneras distintas, como: (a) usar las puntuaciones de una prueba para tomar decisiones importantes, (b) asumir que las pruebas son medidas infalibles del aprovechamiento, y (c) aceptar las pruebas como "buenas" (tal vez porque las construyen compañías comerciales o entidades externas en vez de las maestras) sin evidencia suficiente.

Hay un principio en que las personas dedicadas y estudiosas de la Medición Educativa coincidimos: el uso exclusivo de los resultados de una prueba de aprovechamiento estandarizada, así como de cualquier otro instrumento de medición, para determinar el aprovechamiento académico estudiantil y tomar decisiones importantes es un grave error ético y técnico. Obviamente, depender de un sólo instrumento puede conducir a decisiones incorrectas. Todo instrumento que se diseña para recoger información de las características humanas (físicas o de otra índole) tiene errores. En efecto, este es el primer supuesto de la Teoría Clásica de las Pruebas: la puntuación observada es el resultado de la puntuación "verdadera o real" (la cual es imposible de medir) y el error de la medición. Tampoco ninguna prueba es capaz de medir o representar la complejidad del aprendizaje logrado por un estudiante de manera exacta. Es inapropiado e injusto, pues, depender solamente de las puntuaciones de una prueba para colocar un estudiante en un programa especial, promover o retener una estudiante en un grado y

evaluar la efectividad de un currículo o programa escolar y el desempeño magisterial.

Como he señalado antes, para evaluar el aprovechamiento estudiantil, es necesario contar con información válida, confiable y útil recopilada con distintas técnicas e instrumentos (Medina-Díaz & Verdejo-Carrión, 2000). La ley NCLB también establece que se deben incorporar diferentes instrumentos de *assessment* en los sistemas de *accountability* (Public Law 107-110, Sec.1111(b)(3)(A); Sec. 1111(b)(3)(C)(i-iv),(vi)). Las puntuaciones en las pruebas de aprovechamiento estandarizadas se podrían combinar con las de otras pruebas y tareas que las maestras o la agencia educativa administran al estudiantado. Además, si se van a usar los resultados para tomar decisiones importantes acerca del estudiantado de una escuela, se debe ponderar si son apropiados o no a partir de la documentación técnica disponible de los instrumentos.

Se podía establecer un sistema de *assessment* coordinado y flexible que integre distintas fuentes de información del aprovechamiento estudiantil (e.g., la que proveen las maestras y la de otras fuentes externas). Por supuesto, este sistema debe armonizar con las características particulares de las escuelas, los distintos procesos educativos y las experiencias que tiene el estudiantado. También hay que reconocer que los procedimientos adecuados para combinar las distintas escalas y puntuaciones de los instrumentos que se utilicen, así como los costos que conlleva desarrollarlos y administrarlos pueden ser un obstáculo en su implementación (Koretz & Hamilton, 2006). Jones (2004) y Neill (2004) proponen alternativas para que las escuelas puedan rendir cuentas sobre el cumplimiento de sus responsabilidades académicas, administrativas y sociales considerando múltiples fuentes de información. Estos modelos se podrían considerar, implantar y evaluar en algunas escuelas del país.

Por otra parte, la ideología de que la privatización es sinónimo de calidad también permea el uso de las pruebas de aprovechamiento estandarizadas. Se asume que, por el mero hecho de que las construye una compañía comercial o "alguien de afuera", es un producto de calidad. Sin embargo, una consulta a las críticas de cientos de pruebas de aprovechamiento que se han publicado en los 16 volúmenes del *Mental Measurement Yearbook* demuestra todo lo contrario. Aún cuando las elaboran compañías comerciales reconocidas, hay diferencias en la manera en que estas pruebas cumplen con los estándares de calidad (AERA, APA & NCME, 1999). También es posible que las compa-

ñías cometan errores en la calificación de las pruebas (Neill, 2003b; Henriques, 2003).

Cabe preguntarse, entonces, por qué las pruebas de aprovechamiento estandarizadas se siguen utilizando a pesar de las críticas y limitaciones. Coincido con Lawton, Turner y Roth (1991) en que se debe a dos factores principales: (1) las grandes ganancias de las compañías que construyen las pruebas y venden los servicios y materiales relacionados, y (2) la confianza del público en los datos cuantitativos. Las seis compañías (Harcourt Educational Measurement, CBT McGraw-Hill, Riverside Publishing, Pearson Educational Measurement y Educational Testing Service (ETS), American College Testing) que dominan el mercado de las pruebas de aprovechamiento estandarizadas en los estados y territorios de los Estados Unidos de América aumentaron dramáticamente sus ganancias a partir de la década del noventa (Clarke, Madaus, Horn & Ramos, 2001; Miner, 2004/2005). A menudo, el costo de la prueba incluye diseñarla, construirla, administrarla y calificarla así como interpretar las puntuaciones y producir informes individuales y grupales. La GAO estimó que, entre los años 2002 y 2008, los estados gastarían entre 1.9 billones (si todos usan pruebas de alternativas múltiples) y 5.3 billones de dólares (si combinan ítems de alternativas múltiples con pocas preguntas para elaborar la respuesta) en la implementación de los sistemas de pruebas que exige la ley NCLB (GAO, 2003). En el 2002, el DEPR contrató a la compañía ETS para la construcción y administración de la *Prueba Puertorriqueña de Aprovechamiento Académico* (PPAA) a un costo de $8 millones (véase Tabla 1). Desafortunadamente, la "secretividad" acerca de los contratos, las condiciones y los procedimientos que usan estas compañías en cada estado y territorio no permite la supervisión, la evaluación y la fiscalización de sus operaciones, productos y servicios. Además, tampoco se conocen los criterios que utilizan las agencias educativas para la selección de estas compañías.

Aunque las cifras mencionadas representan sumas alarmantes de dinero, según Haertel (1999) y Linn (2000), es más barato y toma menos tiempo implementar un programa de pruebas que hacer cambios substanciales e importantes en los sistemas y las prácticas escolares (e.g., reducir el tamaño de los grupos, aumentar el tiempo de clases, aumentar la cantidad y el salario de las maestras, proveer y actualizar los libros de texto). Así que, la construcción y administración de estas pruebas es un negocio muy lucrativo para las compañías y ventajoso para los intereses de los políticos. Según Martínez Ramos (2006), tam-

bién encubre el fortalecimiento empresarial como parte de la doctrina neoliberal y de una economía globalizada.

La visibilidad de los resultados, sin lugar a dudas, también ofrece ventajas políticas. A los políticos les conviene reportar en la prensa puntuaciones bajas al principio y luego, mostrar que han aumentado. De esta manera, pueden dar la impresión de que hay ganancias en las puntuaciones, sin que necesariamente haya ocurrido una mejoría "real" en el aprovechamiento estudiantil. Linn, Dunbar, Harnish y Hastings (1982) resumen una serie de factores que pueden inflar o tergiversar los estimados de ganancia en las puntuaciones: (a) la selección y las habilidades intelectuales del estudiantado, (b) los errores de conversión en las escalas, (c) las diferencias en las condiciones de administración, (d) los efectos de la práctica y (e) la enseñanza dirigida a aprobar las pruebas. Ciertamente, en las conferencias de prensa que se citan en Puerto Rico para informar los resultados de las pruebas no se indican ni se preguntan por los procedimientos estadísticos que se aplicaron para calcular las ganancias o reducciones en las puntuaciones.

Además, los políticos, el público y la prensa revisten las puntuaciones de una prueba con un "aura" o "poder" de rigor científico y "objetividad" que les permite hacer inferencias del aprendizaje estudiantil y comparar escuelas, distritos, estados y hasta, países. La atención a los datos numéricos de estas pruebas se basa en una "fe ciega" en la información cuantitativa. También se sostiene en la creencia de que se puede informar o representar el aprendizaje mediante un numeral y no se cuestiona cómo se obtiene. La prueba se convierte en un "instrumento privilegiado" que tiene el "poder extraordinario" de proporcionar información sobre algo que no podemos apreciar a simple vista. De hecho, varios estudios señalan que tanto las maestras como los padres y las madres no entienden cómo se construyen estas pruebas y cómo se obtienen e interpretan las puntuaciones (Apel & Rieche, 2001; Herman & Golan, 1993; Shepard & Bliem, 1995).

No se cuestiona la construcción de la prueba, los estándares curriculares y de ejecución seleccionados, la validez, la confiabilidad, la generalización de los resultados ni las estadísticas derivadas para la interpretación de las puntuaciones (e.g., los porcentajes de proficiencia y los AYP). Además, la publicación de los resultados de las pruebas en la prensa escrita, en los primeros meses del año escolar, genera un interés "general e instantáneo" en la educación pública que incita al público y a los políticos a criticar el sistema educativo y hacer expresiones acerca de los "logros de los estándares y las competencias" (sin

saber que son), así como de las variables que los provocaron. Se formulan, también, inferencias simplistas de *constructs* amplios y complejos como la "proficiencia en Matemática" sin indicar su significado y los elementos específicos de la ejecución del estudiantado que apoyan esa inferencia.

Situación en Puerto Rico

A partir de la década del noventa, el DEPR ha utilizado diferentes pruebas estandarizadas para medir el aprovechamiento académico de los estudiantes: la *Prueba Aprenda*[17], la *Prueba Senda,* las *Prueba Puertorriqueñas de Competencia Escolar* y la *Prueba Puertorriqueña de Aprovechamiento Académico*. La prensa del país ha destacado en sus titulares los resultados de estas pruebas de la siguiente manera: 'Domina el 77 por ciento las destrezas básicas' (Ferré Rangel, 22 de agosto de 1991); 'Devastadores los resultados de la prueba de aptitud' (Millán Pabón, 18 de octubre de 1996); 'Progreso en la educación' (Millán Pabón, 8 de agosto de 1997); 'Flojos' los estudiantes sistema público (Rosario, 11 de septiembre de 2003); 'Medio mundo 'colgado' (Negrón Pérez, 1 de octubre de 2004) y 'Mejora en el aprovechamiento académico' (Roldán Soto, 2 de agosto de 2005). La Tabla 1 presenta algunos de los propósitos y resultados de las pruebas administradas en Puerto Rico.

De la información provista por los funcionarios del DEPR a la prensa del país se desprende que, con las pruebas, se han perseguido propósitos distintos: (a) identificar las destrezas que el estudiantado no domina, (b) detectar fortalezas y deficiencias en el aprendizaje estudiantil, (c) identificar escuelas con pobre rendimiento y (d) servir como un indicador del aprovechamiento académico del sistema público (véase Tabla 1). Al parecer, estos propósitos dependen de la corriente política y legislación de cada momento histórico. En las pruebas que se administraron desde el 2002, no hay una concordancia clara entre los propósitos y los resultados que se destacan. No se ofrece información acerca de la calidad de las pruebas a individuos ni al público general. Tampoco ha estado disponible para la autora de este artículo cuando la ha solicitado.

En el 2003, la PPAA se administró a los estudiantes de tercero, sexto, octavo y undécimo grado en las asignaturas de Español, Matemáticas e Inglés. Esta prueba incluye ítems de alternativas múltiples y las puntuaciones se interpretan en tres niveles de dominio: básico (dominio parcial de las destrezas y conceptos), proficiente (domina la mayor

parte de los conceptos y destrezas) y avanzado (amplio dominio y aplicación de conceptos y destrezas). Según los artículos en la prensa, la prueba se basa en los estándares curriculares de estas asignaturas y se preparó con la participación de 58 maestros del sistema público junto a los directores y los supervisores de los programas académicos y el personal de la compañía ETS (Rodríguez Cotto, 26 de abril de 2003; Negrón Pérez, 26 de abril de 2003). En el año escolar 2003-2004, los resultados de la PPAA indicaron que 42% de los estudiantes de tercero, sexto, octavo y undécimo grado aprobaron la materia de Español, 50% Inglés y 46% Matemática (Negrón Pérez, 1 de octubre de 2004).

Cerca de 300,000 estudiantes de tercero a octavo y undécimo grado contestaron la PPAA en abril del 2005. Los resultados indican que los estudiantes aumentaron su rendimiento en las tres asignaturas incluidas en la prueba (52% en Español, 57% en Matemática y 55% en Inglés) en comparación con los dos años anteriores (Roldán Soto, 28 de enero de 2006; Ruiz Kuilan, 28 de enero de 2006). El DEPR había trazado las siguientes metas en las tres asignaturas: 54% en Matemática,49% en Español y 34% en Inglés. Así que estos resultados superaron lo esperado en Matemática e Inglés. Los estudiantes de 40 escuelas lograron resultados excelentes (100% proficiencia) en las asignaturas de la prueba. Para el año 2010, se espera que 69% del estudiantado logre los niveles de proficiente o avanzado en Matemáticas, 66% en Español y 34% en Inglés.

En el año escolar 2005-2006, 297,000 estudiantes contestaron la PPAA. Sólo 44% de los estudiantes demostraron ser proficientes en Español, 50% en Matemáticas y 50% en Inglés. Como se puede observar, los resultados de ese año reflejan una reducción en la cantidad de estudiantes en los niveles proficiente o avanzado en las tres asignaturas con respecto al año 2005 (Prensa Asociada, 13 de julio de 2006). Cuarenta y seis mil estudiantes del programa de Educación Especial contestaron la prueba, de los cuales sólo una tercera parte mostraron ser proficientes en las tres materias. Sólo 16 escuelas lograron, en algunos de los grados, los niveles de proficiente o avanzado en las tres materias. Tanto en estos resultados como en los anteriores se puede observar que los estudiantes de escuela elemental, especialmente los de tercer grado, obtienen las mejores puntuaciones.

En el año escolar 2006-2007, el DEPR contrató a la compañía *Pearson Educational Measurement* (PEM) para la construcción y administración, corrección y producción de los informes de la PPAA. Esta prueba se administró a 289,313 estudiantes de tercero a octavo y

úndecimo grado en este año. Según los datos publicados en el periódico *El Nuevo Día* (Hernández Cabiya, 31 de mayo de 2007), los porcentajes de estudiantes en estos grados clasificados como proficientes o avanzados en Español fluctuaron entre 41% (octavo grado) y 59% (tercer grado) y en Matemáticas oscilaron entre 42% (séptimo grado) y 78% (tercer grado). En la prueba de Inglés, el porcentaje de estudiantes en la misma categoría fluctuó entre 48% (quinto y séptimo grado) y 56% (tercer y cuatro grado). Si se calcula la media aritmética de los porcentajes de estudiantes en los siete grados clasificados como proficientes o avanzados en las tres asignaturas se obtienen los siguientes resultados: 50% en Español, 55% en Matemáticas y 52% en Inglés. La noticia destaca un aumento en el aprovechamiento estudiantil en comparación con el del año pasado pero que la ejecución es similar a la del estudiantado que contestó la PPAA en el 2005. También señala que se administraron las *Pruebas de Evaluación Alterna* a 1,931 de los 15,990 estudiantes del programa de Educación Especial. Los resultados reflejan un rendimiento mayor que en el año 2006. Para el próximo año escolar, la compañía PEM debe construir y administrar una prueba de Ciencia en la escuela elemental, intermedia y superior.

Lamentablemente, la información publicada en la prensa sobre las pruebas de aprovechamiento que se han administrado y sus resultados es incompleta y pobre, pues no indaga en los estudios acerca de la calidad técnica de la prueba, la definición de aprovechamiento, el método de que utiliza para identificar las escuelas en plan de mejoramiento, la selección del estándar de ejecución para cada año, el trasfondo que sirve de marco para interpretar los niveles de proficiencia y el establecimiento los niveles de dominio (o "metas", como suele llamarle) de las asignaturas examinadas en cada año escolar hasta el 2014, entre otros asuntos. Así que es necesario mayor difusión pública y explicación de los procedimientos que el DEPR utiliza para explicar estos y otros asuntos técnicos de las pruebas. En la página electrónica del DEPR, bajo el título de PPAA, sólo se pueden obtener los resultados de una escuela (si se escribe el nombre) durante los años escolares 2004-2005 y 2005-2006 (http://www.de.gobierno.pr/dePortal/Escuelas/Directorios/PAAIntro.asxp).

Los funcionarios del DEPR subrayan que el contenido está alineado con los estándares de contenido de las asignaturas, según lo exige la ley NCLB, y que están bien construidas. En una conversación telefónica (23 de agosto de 2006), el Sr. Ángel Canales, ex-Director de la Oficina de Evaluación del DEPR, me informó que, en la sección de

Asuntos Federales de la página electrónica, se encontraban los informes de tres estudios acerca de la alineación de las pruebas con los estándares curriculares de Matemática, Español e Inglés. En los estudios de Norman Webb (2005 a,b,c), entre seis y ocho jueces evaluaron la alineación de los ítems de la PPAA con los estándares curriculares y sus respectivos objetivos. Los cuatro criterios que utilizaron fueron: (1) *categorical concurrence* (si los items que miden el contenido del estándar); (2) *depth-of-knowledge consistency* (si las demandas cognitivas de conocimiento de los ítems corresponden a lo que el estándar espera de los estudiantes); (3) *range-of-knowledge correspondence* (si la amplitud del conocimiento que se espera del estudiantado en un estándar corresponde a lo que se necesita para contestar los ítems) y (4) *balance of representation* (si hay una distribución similar entre los ítems que representan los objetivos bajo un estándar).

Los resultados indican que hay cierta alineación entre los estándares curriculares y las pruebas pero que necesitan mejorar sustancialmente, bien sea remplazando o modificando gran cantidad de ítems. La mayoría de los ítems de las pruebas no cumplen con los cuatro criterios de alineación ni representan objetivos de niveles cognoscitivos altos. Las pruebas de Matemáticas de cuarto y undécimo grado; y de Español en cuarto, quinto, séptimo, octavo y undécimo grado son las peores "alineadas" con los estándares. Las pruebas de Español que se diseñaron para tercero y sexto grado cumplen con los criterios de alineación en sólo un estándar. Aunque los ítems de todas las pruebas de Inglés representan tres de los cuatro estándares, hay mejor alineación en las del nivel elemental. Weeb (2005 a,b, c) encontró también que los ítems de las tres pruebas no miden adecuadamente la amplitud del conocimiento que se espera de los estudiantes en los estándares curriculares. De acuerdo con los resultados de este análisis no se puede concluir o proclamar que las PPAA "están alienadas" a los estándares curriculares de las asignaturas.

Hasta ahora no hay otra información acerca del cumplimiento de uno de los requisitos de calidad técnica (i.e., validez de contenido) de estas pruebas (AERA, APA & NCME, 1999). En la página electrónica de la compañía ETS (http://www.ets.org) tampoco aparece información relacionada con la PPAA. Por lo tanto, sólo conocemos los datos que los funcionarios del DEPR presentan en las conferencias de prensa. No obstante, parece que las escuelas y las maestras reciben más información de los resultados de las pruebas que la que he reseñado (entrevista al Sr. Ángel Canales, 24 de enero de 2007). Según este funcionario

del DEPR, cada estudiante y sus padres o madres reciben información sobre el desempeño. De hecho, la ley NCLB dispone que se generen reportes individuales de cada estudiante con información descriptiva, diagnóstica y de la interpretación de las puntuaciones, de manera que los padres, las madres, las maestras y las directoras escolares puedan entender y atender las necesidades académicas especificas de los estudiantes (Public Law 107-110, Sec. 1111(b)(3)(C)(xii)).

Una consecuencia directa del bajo rendimiento de los estudiantes en las pruebas es la clasificación de sus escuelas como "en plan de mejoramiento", según lo dispone la ley. Se utilizan las puntuaciones en las pruebas así como otros datos de la escuela (e.g., matrícula, nivel de retención y promoción) para calcular el AYP. Cabe señalar que la posibilidad que tiene una escuela de aumentar el AYP, paulatinamente, depende de dónde comenzó. Las escuelas que son más proficientes tienen una mayor probabilidad de lograr el APY en un lapso de dos años, que las que no lo son (Linn, 2003). Además, el DEPR ha delegado el cómputo del AYP a varias compañías o personas en los últimos cinco años (Entrevista con Ángel Canales, 27 de enero de 2007), lo que podría influir en las diferencias y posibles errores en los cálculos a través de los años.

De acuerdo con los datos del DEPR, la cantidad de escuelas en el país identificadas con "plan de mejoramiento" ha aumentado desde el 2002. En los años 2005-2006, la cantidad de escuelas en plan de mejoramiento aumentó a 674 (44%): 251 escuelas comenzaban en el plan de mejoramiento, 323 estaban en el segundo año, 68 en el tercero, 24 en el cuarto y ocho en el quinto. En la página electrónica del DEPR hay una lista de 701 escuelas que estaban o comenzaron en el plan de mejoramiento en los últimos dos años escolares. Aunque la ley NCLB establece sanciones —como el cierre de las escuelas— para las que llegan al quinto de año en el plan de mejoramiento, el Subsecretario de Asuntos Académicos, Waldo Torres, indicó que a estas escuelas "se les iba a ofrecer ayuda adicional" (Roldán Soto, 2 de agosto de 2005).

A la inversa, la prensa reseñó que ocho escuelas que llevaban seis años en "plan de mejoramiento" no han podido cumplir con lo que dispone esta ley porque no había recibido el apoyo prometido (Roldán Soto, 29 de enero de 2007). La periodista señala que el DEPR "logró que el gobierno federal no les someta las severas sanciones" que contempla la ley "pero ha hecho muy poco para sacarlas permanentemente de su difícil situación" (Roldán Soto, 29 de enero de 2007, p. 7). El Subsecretario de Asuntos Académicos, profesor Waldo Torres,

reconoció que el DEPR no hizo nada para atender la situación de estas escuelas y que la agencia elaboró un plan de acción que contempla proveerle apoyo técnico y atender los problemas de las escuelas.

Ante esta situación patética, el estudio de Vázquez Pérez (2006) asoma alguna esperanza para estas escuelas. El autor documenta las estrategias y prácticas que una escuela elemental e intermedia llevó a cabo, por espacio de dos años, para lograr superar la clasificación de "estar en plan de mejoramiento" y con muy poca ayuda del DEPR. La comunidad escolar creó lo que denominaron la "cultura de la prueba puertorriqueña". Esta "cultura" se apoya en la divulgación de la ley NCLB y la responsabilidad compartida de toda la comunidad escolar, la discusión y el análisis de los resultados de las pruebas, el énfasis en la importancia de las pruebas y la práctica de estrategias para la enseñanzas en Inglés, Matemáticas y Español utilizando los estándares curriculares y el contenido de las pruebas. Estas acciones concuerdan con las que otros autores han señalado.

Conclusiones

El uso de las pruebas de aprovechamiento estandarizadas como instrumento para controlar y determinar la calidad del sistemas escolar público han tomado mayor realce y poder en los últimos seis años en los Estados Unidos de América y Puerto Rico. Debido a las disposiciones de la ley NCLB y las presiones de los sectores políticos y económicos, los resultados de estas pruebas tienen repercusiones directas y graves para las agencias educativas, las escuelas, el magisterio y el estudiantado. La reducción paulatina de fondos federales es la penalidad principal (Serrano, 18 de agosto de 2006). Esto coloca a las agencias educativas y escuelas en una condición de alto riesgo ya que conlleva alguna sanción o penalidad si no logran el rendimiento académico que se espera. Ante la falta de recursos y peritaje técnico, casi todas las agencias educativas de los estados y territorios han confiado a compañías comerciales la construcción de estas pruebas. Por consiguiente, gastan millones de dólares en los contratos de estas empresas en vez de invertirlos en los sistemas y las comunidades escolares. Por esto, Neill (2003a) y Kohn (2004) subrayan que, detrás de esta ley, hay una agenda de hostilidad y privatización de la educación pública.

Lo más lamentable del uso de las puntuaciones de estas pruebas es que hay pocas investigaciones que apoyen los posibles beneficios, los cambios y el impacto en la educación que la ley promulga (Vázquez Pérez y Bonilla Rodríguez, 2006). La evidencia disponible destaca

los efectos negativos en las escuelas, el magisterio y el estudiantado. Así que, contrario a fomentar una "verdadera" reforma educativa y crear oportunidades para que el estudiantado logre las expectativas que resaltan los estándares, demuestren aprendizajes complejos y sus talentos individuales, las pruebas de aprovechamiento han provocado la atención y dedicación de tiempo substancial a la enseñanza del contenido de las mismas y a la preparación para contestarlas así como la proliferación de prácticas deshonestas (Haertel, 1999; Herman & Golan, 1993; Darling-Hammond & Rustique-Forrester, 2005; Vázquez Pérez, 2006). Esto, por supuesto, produce una inflación en las puntuaciones que conduce a inferencias desacertadas acerca del dominio que tiene el estudiantado de los estándares de contenido representados en las pruebas (Koretz, 2005).

Por las razones expuestas en este artículo, se desaprueba la práctica de la dependencia exclusiva de estas pruebas para tomar decisiones que tienen consecuencias adversas. De ninguna manera, se ha demostrado que el mero uso de las pruebas de aprovechamiento estandarizadas conduce a mejorar el aprendizaje estudiantil y la calidad de los sistemas educativos. La administración repetida de las pruebas tampoco asegura que se van a lograr los estándares y aprendizajes complejos que se esperan. Todo lo contrario, los resultados de los estudios de Wolf y Smith (1995), Brown y Walberg (1993), Sadker y Zittleman (2004) revelan que el estudiantado se puede sentir desmotivado y desinteresado en contestar las pruebas, si estas no tienen consecuencias directas en su aprovechamiento escolar. Más aún, cabe la posibilidad de que exista alguna relación entre la incidencia en la deserción escolar y los resultados de las pruebas de aprovechamiento en los estados (Darling-Hammond, 2004).

La "fe ciega" en las puntuaciones de estas pruebas también imposibilita a los políticos, los funcionarios de las agencias educativas y al público en general, apreciar sus limitaciones como un instrumento de medición del aprovechamiento escolar. Existe la creencia de que las puntuaciones de las pruebas son claras y se entienden porque son numerales. Se presume que las puntuaciones de las pruebas son infalibles y se confía en su "objetividad". Como he señalado antes: las pruebas son instrumentos de medición imperfectos. Sólo permiten una visión limitada, una vez al año, del posible conocimiento adquirido en un puñado de ítems. Además, se debe tener cuidado con la noción de "adquisición y dominio de la materia o de los estándares" (como sostienen algunos de los funcionarios del DEPR en la prensa) cuando no

hay evidencia suficiente acerca de la validez de las puntuaciones que le sirva de apoyo. Así que, no se puede presumir, sin la debida evidencia, que la PPAA u otra prueba similar "mide" los estándares de contenido de una asignatura por grado. Es necesario demostrar teórica, lógica y empíricamente que la prueba esta "alineada" con los mismos. La evaluación de la alineación de la PPAA con los estándares revela su incumplimiento (Webb, 2005 a, b, c).

Tampoco hay evidencia suficiente acerca del cumplimiento con los estándares de calidad técnica (AERA, APA & NCME, 1999) y de la contribución que estas pruebas pueden hacer al mejoramiento de la enseñanza de los estándares de contenido en las distintas escuelas del país. Hay mucha retórica política y poca evidencia empírica y técnica que la apoye. Anticipo que, en los próximos años, se continuarán usando, de manera errada, los resultados de las pruebas para fines políticos. Es urgente, pues, que el DEPR y las compañías que contrata para construir las pruebas de aprovechamiento publiquen información acerca de la validez de sus puntuaciones e inferencias así como sus consecuencias (conocida como *consequential validity*). La validez en las pruebas y otros instrumentos de *assessment* radica en la integración de evidencia de diversas fuentes (e.g., contenido, proceso cognitivos, estructura interna y consecuencias) que apoyen las interpretaciones, inferencias y decisiones que se realizan a partir de las puntuaciones (AERA, APA & NCME, 1999). La descripción de la prueba y las especificaciones técnicas deben incluir esta información y la misma debe estar accesible a las personas interesadas. Por las implicaciones políticas, sociales y económicas que tienen las puntuaciones de las pruebas, esto es lo menos que el DEPR y las compañías que contrata deben hacer. Urge, también, que un organismo o comité independiente evalúe el proceso mediante el cual el DEPR está implementando el sistema de *accountability,* así como las prácticas en las escuelas y los distritos para cumplir con los requisitos de la ley NCLB. Los *Standards for Educational Accountabilty Systems* (Baker, Linn, Herman & Koretz, 2002) pueden servir de referencia en esta evaluación.

En fin, las pruebas de aprovechamiento estandarizadas que se administran en los Estados Unidos de América y Puerto Rico presentan una serie de limitaciones que provocan cuestionar la validez y uso de los resultados. La alineación entre los estándares de contenido, las pruebas, el currículo y la enseñanza en las escuelas es crucial para reclamar la validez y apoyar el sistema de *accountability* que la ley NCLB impone. La evidencia disponible es contraria. Además,

los funcionarios de las agencias educativas y los políticos han usado e interpretado de manera inadecuada los resultados de las pruebas para penalizar a las agencias educativas y comunidades escolares por el bajo aprovechamiento estudiantil. Han usado las pruebas como pretexto para ocultar las verdaderas causas y factores que provocan las dificultades y deficiencias del sistema escolar público. Esto revela que, para los intereses políticos y económicos, resulta más beneficioso aplicar las pruebas de aprovechamiento que atender y solucionar los diversos problemas que confrontan las escuelas. El énfasis debe ser en entender por qué unas escuelas logran los resultados esperados y otras no y auscultar los factores que influyen en la ejecución del estudiantado. La investigación educativa es, sin lugar a dudas, un recurso para buscar esas explicaciones e identificar posibles soluciones. ¿Qué estamos esperando?

Tabla 1. Pruebas de aprovechamiento estandarizadoas administradas en Puerto Rico 1990-2005*

PRUEBA/AÑO ESCOLAR	PROPÓSITOS	RESULTADOS
Aprenda/1990-1991 Costo anual: $ 1.4 millones	"Los resultados de las pruebas facilitará la educación más individualizada; permitirá a los maestros tener un cuadro detallado del dominio de destrezas de sus estudiantes; ayudará a los padres a tener un conocimiento más completo del dominio de las destrezas de sus hijos y facilita a los estudiantes entender mejor sus fortalezas y debilidades en el proceso de aprendizaje". Los resultados de las pruebas también han servido para detectar cuáles son las escuelas con menos rendimiento de manera que el Departamento pueda atenderlas prioritariamente. "En el Departamento vamos a ofrecer a las escuelas de baja ejecución nuestro apoyo... Queremos que estos resultados generen un proceso de autoevaluación" (Celeste Bénitez, citada por Ferré Rangel, 22 de agosto de 1991, pág.28)	De los 664,000 estudiantes en todos los grados del sistema público: 77.22% dominó las destrezas de Inglés, 77.36% las destrezas de Español, y 77.43% las de Matemáticas (Ferré Rangel, 22 de agosto de 1991, pág.28)
Pruebas Puertorriqueñas de Competencia Escolar/1995-1996 Costo anual: Entre $ 700,000 y $ 2.9 millones	Identificar las deficiencias en el aprendizaje de los estudiantes en las materias de Español, Inglés, Matemáticas, Ciencias y Estudios Sociales (Millán Pabón, 18 de octubre de 1996). "Esta prueba se administró de forma experimental para ir examinando el contenido, ver si se atempera a lo que dan los maestros y examinar si los enfoques que traemos en la reforma educativa estaban contemplados en la prueba" (Isidra Albino, citada por Millán Pabón, 18 de octubre de 1996).	De 436,737 estudiantes de tercero a duodécimo: 38% aprobó la materia de Español, 19% Inglés, 18% Estudios Sociales, 26% Matemáticas (Millán Pabón, 19 de octubre de 1996, pág. 12),

PRUEBA/AÑO ESCOLAR	PROPÓSITOS	RESULTADOS
Prueba Puertorriqueña de Aprovechamiento Académico/2002-2003 Costo: $ 8 millones	"Le proveera al país indicadores sobre la realidad académica del sistema público de enseñanza" (César Rey, citado por Negrón Pérez, 26 de abril de 2003). "Estas pruebas son un complemento de otros indicadores de aprovechamiento académico como los informes de notas y los comentarios de los maestros. La suma de estos indicadores mide el progreso total del estudiante" (César Rey, citado por Rosario, 11 de septiembre de 2003, pág.11).	De los 170,000 estudiantes de tercero, sexto, octavo y undécimo grado que contestaron las pruebas: 48% aprobaron la materia de Español, 46% en Matemáticas y 50% en Inglés. (Rosario, 11 de septiembre de 2003, pág. 11; Rodríguez Cotto, 11 de septiembre de 2003, pág. 28).
Prueba Puertorriqueña de Aprovechamiento Académico/2003-2004 Costo: No hay información disponible.	"Este año, las pruebas miden por competencia, por destreza. Vamos a poder identificar las destrezas que el estudiante no domina, para dar a los maestros adiestramiento" (Brunilda Martínez, citada por Negrón Perez, 1 de octubre de 2004, pág. 3).	42% de los estudiantes de tercero, sexto, octavo y undécimo grado aprobaron la materia de Español, 50% Inglés y 46% Matemática (Negrón Pérez, 1 de octubre de 2004, pág. 3).
Prueba Puertorriqueña de Aprovechamiento Académico/2004-2005 Costo: No hay información disponible.	No hay información disponible.	De 300,000 estudiantes de tercero a octavo y undécimo grado: 52% demostraron ser proficientes en Español, 57% en Matemática y 55% en Inglés (Roldán Soto, 28 de enero de 2006; Ruiz Kuilan, 28 de enero de 2006, pág. 15).

PRUEBA/AÑO ESCOLAR	PROPÓSITOS	RESULTADOS
Prueba Puertorriqueña de Aprovechamiento Académico/2005-2006 Costo: No hay información disponible.	No hay información disponible.	De los 297,000 estudiantes en 1,469 escuelas: 44% de los estudiantes demostraron ser proficientes en la materia de Español, 50% en Matemáticas y 50% en Inglés 46,000 (16%) estudiantes del programa de Educación Especial contestaron la prueba (Prensa Asociada, 13 de julio de 2006, pág. 16).
Prueba Puertorriqueña de Aprovechamiento Académico/2006-2007 Costo: No hay información disponible.	"Se mide el progreso anual del desempeño de cada escuela, y por consiguiente, el aprovechamiento académico de los estudiantes" (Rafael Aragunde, citado por Resto Vélez, 3 de abril de 2007, pág. 7).	De 289, 313 estudiantes de tercero a octavo y undécimo grado 50% se clasificaron como proficientes o avanzados en las pruebas de Español, 55% en Matemáticas y 52% en Inglés**. 1.931 (12%) estudiantes del programa de Educación Especial contestaron las *Pruebas de Evaluación Alterna* y mostraron un mayor rendimiento que en los del año 2006 (Hernández Cabiya, 31 de mayo de 2007, pág. 8).

Notas: *No se recomienda comparar los resultados de estas pruebas por las siguientes razones: (1) hay diferencias en el propósito, contenido e interpretación de los resultados de las pruebas; (2) los estudiantes de cada grado que contestan las pruebas son diferentes año tras año; (3) existen diferencias en las técnicas de enseñanza y evaluación así como en los materiales didácticos que utilizan las maestras a lo largo del tiempo; (4) hay distintos énfasis y modalidades en la enseñanza del contenido que incluyen las pruebas; (5) la representación del contenido y de los diferentes niveles cognitivos en los ítems fluctúa entre las pruebas, aún cuando la misma compañía construya la prueba y lleve el mismo título; y (6) no hay suficiente evidencia empírica que sustente la calibración y equivalencia de las puntuaciones en las pruebas que se administraron en dos o tres años consecutivos, aunque sean de la misma compañía y las pruebas tengan el mismo título. ** Estos porcentajes representan la media aritmética de los porcentajes reportados en cada una de las pruebas administradas en los siete grados escolares.

REFERENCIAS

Abedi, J. y Dietel, R. (Winter, 2004). Challenges in the No Child Left Behind Act for English Language Learners. *Policy Brief 7*, National Center for Research on Evaluation, Standards, and Student Testing, Univesity of California, Los Angeles.

American Educational Research Association, American Psychological Association y National Council of Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Apel, J. y Rieche, B. (2001). *Las pruebas en el aula: Aprendizaje y evaluación.* Buenos Aires, Argentina: Aique Grupo Editor.

Baker, E. L., Linn, R. L., Herman, J. L. y Koretz, D. (Winter, 2002). Standards for educational accountability systems. *CRESST Policy Brief 5*, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Bhola,D.S., Impara, J. C. y Buckdenhal, C.W. (2003). Aligning tests with states' content standards: Methods and issues.*Educational Measurement: Issues and Practice*, 22(3), 21-29.

Berliner, D. C. (2006). Our impoverished view of educational research. *Teachers College Record, 108* (6), 949-995.

Brown, S. M. y Walberg, H. J. (1993). Motivational effects of test scores of elementary students. *Journal of Educational Research*, 86, 133-136.

Caquías Cruz, S. (7 de mayo de 2005). Bajo fuego unas pruebas escolares, *El Nuevo Día*, 14.

Carnevale, A. P. y Kimmel, E. W. (1997). *A national test: Balancing policy and technical issues.* Princeton, NJ: Educational Testing Service.

Choi, K., Goldschmidt, P. y Yamashiro,K. (2005). Exploring models of school performance: From theory to practice. En En L. K. Herman y E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (104th Yearbook of the National Society for the Study of Education, pp.119-146). Malden, Mass: Blackwell.

Cizek, G. J. (Spring, 2001). Cheating to the test. *Education Matters,* 41-47.

Chester, M. D.(2005). Making valid and consistent inferences about school effectiveness from multiple measures.*Educational Measurement: Issues and Practice*, *24* (4), 40-52.

Clarke, M., Madaus, G., Horn, C. y Ramos, M. (Abril, 2001). *The marketplace for educational Testing, 2*(3), National Board of Educational Testing and Public Policy, Boston College. Recuperado de http://www.bc.edu/research/nbetpp/publications/v2n3.html en 8/24/2006.

Clune, W. H. (2001). Toward a theory of standard-based reform: The case of nine NSF Statewide System Initiatives. En S.H. Fuhrman (Ed.), *From the capitol to the classroom: Standards-based reform in the states* (One hundredth yearbook of the National Society for the Study of Education, pp. 13-38). Chicago, ILL: University of Chicago Press.

Comisión para el Desarrollo de los Estándares Curriculares y de "Assessment" para la Ciencia y Matemática Escolar en Puerto Rico (Marzo, 1996). *Estándares curriculares para la matemática escolar en Puerto Rico*. San Juan, PR: Consejo General de Educación, Centro de Recursos para la Ciencia e Ingeniería y Departamento de Educación de Puerto Rico.

Darling-Hammond, L. (2004). From "Separate but Equal" to "No Child Left Behind": The collection of new standards and old inequalities. En D. Meier y G. Wood (Eds.), *Many children left behind: How the No Child Left Behind is damaging our children and our schools* (pp.3-32). Boston, Mass: Beacon Press.

Darling-Hammond, L. y Rustique-Forrester, E. (2005). The consequences of student testing for teaching and teaching quality. En L. K. Herman y E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (104th Yearbook of the National Society for the Study of Education, pp. 289-319). Malden, Mass: Blackwell.

Delgado Santa Gadea, K (1996). *Evaluación y calidad en la educación: Nuevos aportes, procesos y resultados*. Santafé de Bogotá, Colombia: Cooperativa Editorial Magisterio.

Dobbs, M. (June 7, 2004). Poor schools sue for funding. *Washington Post,* A13.

Ferré Rangel, L. A. (22 de agosto de 1991). Domina el 77 por ciento las destrezas básicas*, El Nuevo Día*, 28.

Ferrara, S. y DeMauro, (2006). Standardized assessment of individual achievement in K-12. En R. L. Brennan (Ed.), *Educational Measurement* (4th.ed., pp. 579-621). Westport, CT: American Council on Education & Praeger.

Golberg, M. (2005). Test mess 2: Are we doing better a year later? *Phi Delta Kappan*, *86* (5), 389-395.

Goldschmidt, P. y Choi, K. (Spring, 2007). The practical benefits of growth models for accountability and the limitations under NCLB.*CESST Policy Brief 9*, National Center for Research on Evaluation, Standards, and Student Testing, Univesity of California, Los Angeles.

Goverment Accounting Office (Mayo 8, 2003). *Title I: Characterístics of tests will influence expenses; Information sharing may help states realice efficiencies*. GAO-03-389. Recuperado de http://www.gao.gov/cgi-bin/getrpt?GAO-03-389 el 8/21/2006.

Goverment Accounting Office (Diciembre 10, 2004). *No Child Left Behind Act: Education needs to provide additional technical assisstance and conduct implementation studies for school choice provision.* GAO-0507. Resumen

recuperado de http://www.gao.gov/docsearch/abstract.php?rptno=GAO-057 el en 8/21/2006.

Goverment Accounting Office (Julio 26, 2006). No Child Left Behind Act: Assistance from education could help states better measure progress of students with limited english proficiency. GAO-06-815. Resumen recuperado de http://www.gao.gov/docsearch/abstract.php?rptno+GAO-06-815 en 8/21/2006.

Haladyna, T. M. y Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Testing: Issues and Practice*, *23* (1), 17-27.

Hamilton, L. y Stecher, B. (2004). Responding effectively to test-based accountability. *Phi Delta Kappan*, *85*(8), 578-583.

Harrighton-Lueker, D. (Diciembre, 2000). When educators cheat. *The School Administrator*, 32-39.

Haertel, E. H. (1999). Performance assessment and educational reform. *Phi Delta Kappan*, *80*(9), 662-669.

Haertel, E. H. y Herman,J. L. (2005). A historical perspective on validity arguments for accountability testing. En L. K. Herman y E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (104th Yearbook of the National Society for the Study of Education, pp. 1-34). Malden, Mass: Blackwell.

Henriques, D. B. (September 2, 2003). Rising demands for testing push limits of its accuracy. *New York Times*, A-1, A-16.

Herman, J. L. y Golan, S. (Winter, 1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice,* 20-25, 41.

Hernández Cabiya, Y. (31 de mayo de 2007). Mejora la respuesta estudiantil. *El Nuevo Día*, p.8

Hoover, H. D. (2003). Some common misconceptions about tests and testing. *Educational Measurement: Issues and Practice*, 22(1), 5-14.

Hughes, S. y Bailey, J. (2001/2002). What students think about high-stakes testing. *Educational Leadership*, *59*(4), 74-76.

Joint Committee on Testing Practices (2004).*Code of fair testing practices in education.* Washington, DC: Autor. Recuperado de http://www.apa.prg/science/jctpweb.html en 8/21/2006.

Jones, K. (2004). A balanced school accountability model: An alternative to high-stakes testing. *Phi Delta Kappan*, *85*(8),584-590.

Jones, M. G., Jones, B. D., Hardin, B., Chapman, L., Yarbrough, T. y Davis, M. (1999). The impact of high-stakes testing on teachers and students in North Carolina. *Phi Delta Kappan*, *81*(3), 199-203.

Kohn, A. (2004). Test today, privatize tomorrow: Using accountability to 'reform' public schools to death. *Phi Delta Kappan*, *85*(8), 569-577.

Koretz, D. M. y Hamilton, L.S. (2006). Testing for accountability in K-12. En R.L. Brennan, (Ed.), *Educational Measurement* (4th.ed., pp. 531-578). Westport, CT: American Council on Education & Praeger.

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. En L. K. Herman y E. H. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (104th Yearbook of the National Society for the Study of Education, pp. 99-118). Malden, Mass: Blackwell.

Linn, R. L. (Summer, 2005). Fixing the NCLB accountability system. *CRESST Policy Brief 8*, Policy Brief of the National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles.

Linn, R. L. (Winter, 2003). Requirements for measuring adequate yearly progress. *Policy Brief 6*. National Center for Research on Evaluation, Standards, and Student Testing, Univesity of California, Los Angeles.

Linn. R. L. (2000). Assessments and accontability. *Educational Researcher*, 4-16.

Linn, R. L., Baker, E. L y Herman, J. L. (Fall, 2005). Chickens come home to roost. *CRESST Line*, 1,3, 7-8.

Linn, R. L., Dunbar, S. B., Harnish, D. L. y Hastings, C.N. (1982). The validity of the Title I evaluation and reporting system. En E. R. House, S. Mathison, J. Pearsol, y H. Preskill (Eds.), *Evaluation studies review annual* (Vol. 7, pp.427-442). Beverly Hills, CA: Sage Publications.

Martínez Ramos, L. (2006). "No Child Left Behind" o la seducción del discurso. *Pedagogía*, *39*(1), 58-79.

Medina-Díaz, M. del R. (1998). Review of the Aprenda: La prueba de logros en español. En J. C. Impara y B. S. Plake (Eds), *The Thirteenth Mental Measurements Yearbook* (pp.40-42). Lincoln, NE: The Buros Institute of Mental Measurements.

Medina, M. del R. (Marzo-Abril, 1998). Estándares: Concepto y realidad, *Boletín informativo de la Asociación de Maestros de Matemáticas*, 4-5.

Medina, M. del R. (Noviembre, 1992). Más allá de los estándares, *Boletín Informativo de la Asociación Puertorriqueña de Maestros de Matemáticas,* 6.

Medina Díaz, M del R. y Verdejo Carrión, A.L. (2006). ¿Evaluación o avalúo? *INEVA en acción* (http://ineva.uprrp.edu), *2*(4). 1-5

Medina-Díaz, M. y Verdejo-Carrión, A.L. (2000). *Evaluación del aprendizaje estudiantil.* San Juan, PR: Isla Negra Editores.

Meier, D. y Wood, G. (Eds.), *Many children left behind: How the No Child Left Behind is damaging our children and our schools*. Boston, Mass: Beacon Press.

Millán Pabón, C. (18 de octubre de 1996). Devastadores los resultados de prueba de aptitud, *El Nuevo Día*, 10.

Millán Pabón, C. (8 de agosto de 1997). Progreso en la educación, *El Nuevo Día*, 14.

Miner, B. (Winter, 2004/2005). Testing companies mine gold. *Rethinking schools*, *19*(2), 1-6

Nathan, L. (2002). The human face of the high-stakes testing story. *Phi Delta Kappan, 83*(8), 595-600.

National Commision on Excellence in Education (1983). *At nation at risk*. Washington, DC: U.S. Printing Office.

National Council of Teachers of Mathematics (1989). *Curriculum and Evaluation Standards for School Mathematics*. Reston, VA: Autor.

Negrón Pérez, I. (26 de abril de 2003). Nuevas pruebas de aprovechamiento académico en DE. *El Vocero de Puerto Rico*.

Negrón Pérez, I. (1 de octubre de 2004). Medio mundo 'colgado', *El Vocero de Puerto Rico*, 3.

Neill, M. (2004). Leaving no child behind: Overhauling NCLB. En Meier, D. y G. Wood (Eds.), *Many children left behind: How the No Child Left Behind is damaging our children and our schools* (pp.101-119). Boston, Mass: Beacon Press.

Neill, M. (2003a). Leaving children behind: How no child left behind will fail our children. *Phi Delta Kappan*, *85*(3), 225-228.

Neill, M. (February, 2003b). The dangerous of testing. *Educational Leadership*, 43-46.

Olson, L. (Febrero 1, 2006). States vie to be part of NCLB 'Growth' Pilot. *Education Week*, *25*(21), 24-26.

Paris, S. G., Lawton, T. A., Turner, J. C. y Roth, J. L. (Junio-Julio, 1991). A developmental perspective on standardized achievement testing. *Educational Researcher*, 12-20.

Pedulla, J. J., Abrams, L.M., Madaus, G. F., Rusell, M.K., Ramos, M.A. y Miao, J. (Marzo, 2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teschers. Nation*al Board of Educational Testing and Public Policy. Boston College.

Popham, J. W. (1996). Why standardized tests don't measure educational quality? *Educational Leadership*, *56*(6), 8-15.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*, 3-14.

Prensa Asociada (13 de Julio de 2006). Según pruebas de Educación sólo la mitad de los alumnos da el grado. *Primera Hora*, 16.

Resnick, L.B., Rothman, R., Slattery, J. B. y Vranek, J. L. (2003-2004). Benchmarking and alignment of standards and testing. *Educational Assessment*, *9* , 1-27.

Resto Vélez, M. D. (3 de abril de 2007). Insiste en efectividad de pruebas, *El Vocero*, 7.

Rivera Sánchez, M. (30 de marzo de 2007). Se 'cuelga' educación pública, *El Vocero*, 12

Rodríguez Cotto, S.D. (2003). Novedosos exámenes de aprovechamiento, *El Nuevo Día*, 14.

Roldán Soto, C. (29 de enero de 2006). En el olvido ocho planteles rezagados. *El Nuevo Día*, 6-7.

Roldán Soto, C. (2 de agosto de 2005). Mejora el aprovechamiento académico, *El Nuevo Día*, 12.

Rosario, I. Y. (10 de mayo de 2004). Investigan prueba en escuela Aguada. *El Vocero de Puerto Rico*, 18.

Rosario, I. Y. (11 de septiembre de 2003). Flojos los estudiantes sistema público. *El Vocero de Puerto Rico*, 11.

Ruiz Kuilan, G. (28 de enero de 2006). Estrategia académica con resultados de excelencia. *El Nuevo Día*, 15.

Sadker, D. y Zittleman, K. (2004). Test anxiety: Are students failing tests or are the test failing students?. *Phi Delta Kappan*, *85*(10), 740-744,751.

Serrano, O. J. (18 de agosto de 2006). Por "Ningún niño rezagado" Insuficiencias del DE hacen peligrar fondos. *Primera hora*, 18.

Shepard, L. A. y Bliem, C. L. (Noviembre, 1995). Parents' thinking about standardized tests and performance assessments. *Educational Researcher,* 25-32.

Smith, M. L. (Junio-Julio, 1991). Put to the test: The effects of external testing on teachers. *Educational Researcher*, 8-11.

Smith, M. L. y Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and practice, 10*, 7-11.

Townsend, B. L. (2002). "Testing while black": Standard-based school reform and african american learners. *Remedial and Special Education*, 23(4), 222-230.

Tucker, M. S. & Toch, T. (2004). The secret to making NCLB work? More bureacrats. P*hi Delta Kappan, 86*(1), 28-33.

Vázquez Pérez, J. P. (2006). *Estrategias exitosas utilizadas por una escuela pública para salir del plan de mejoramiento.* Tesis de maestría inédita. Universidad de Puerto Rico, Recinto de Río Piedras.

Vázquez Pérez, J. P. y Bonilla Rodríguez (2006). Necesidad del estudio del impacto de la Ley "No Child Left Behind".*Pedagogía*, *39*(1), 29-57.

Wallis, C. y Steptoe, S. (June 4, 2007). How to fix No Child Left Behind, *Time*, *169*(23), 34-41.

Webb, N. E. (2006). Identifying content for student achievement tests. En S.M. Downing y T.M. Haladyna (Eds.), *Handbook of test development,* (pp. 155-180). Mahwah, NJ: Lawrence Erlbaum.

Webb, N. (Diciembre 19, 2005a). *Alignment Analysis of Spanish Standards and Assessments, Puerto Rico, Grades 3,4,5,6,7,8 and 11.* [Draft Report].

Webb, N. (Diciembre 20, 2005b). *Alignment Analysis of Mathematics Standards and Assessments, Puerto Rico, Grades 3,4,5,6,7,8 and 11.* [Draft Report].

Webb, N. (Diciembre 27, 2005c). *Alignment Analysis of Language Arts Standards and Assessments, Puerto Rico, Grades 3,4,5,6,7,8 and 11.* [Draft Report].

Wolf, L. F. y Smith, J.K. (1995). The consequences of consequence: Motivation, anxiety and test performance. *Applied* Measurement in Education, 8, 227-242.

NOTAS

1 Instrumento se refiere al medio o aparato particular que se utiliza para aplicar una técnica o un procedimiento.

2 La noción de calidad que subyace a este discurso "sirve para designar metas, hábitos y capacidades que pueden someterse a una medición objetiva" (Delgado Santa Gadea, 1996, pág. 33). Así, la medición revela el grado en que los medios y recursos educativos conducen hacia las metas y los objetivos. La "calidad de la educación" se identifica, pues, con resultados (i.e., aprovechamiento o rendimiento académico).

3 Desde el año 1988, el ***National Assessment of Educational Progress*** (NEAP) reporta el aprovechamiento académico de los estudiantes de cuarto, octavo y duodécimo y de los distintos grupos que asisten a las escuelas públicas y privadas de los Estados Unidos de América. También se le conoce como el ***Nation's Report Card***". Consiste de una batería de pruebas en varias asignaturas (Lectura, Matemáticas, Ciencia, Escritura, Historia de los Estados Unidos de América, Civismo, Geografía y Artes) que se administran, periódicamente, a una muestra de estos estudiantes. Estas pruebas siguen las especificaciones que desarrolla el ***National Assessment Governing Board***. El Congreso de los Estados Unidos de América nombra a los miembros de este comité (gobernadores, legisladores, oficiales escolares, educadores, representantes de empresas y del interés público). Para mayor información consulte la página electrónica http://nces.ed.gov/nationsreportcard/

4 Los estados utilizan distintas maneras de nombrar estos estándares (e.g., estándares curriculares, metas, expectativas académicas, guías curriculares, resultados del aprendizaje, marco curricular).

5 ***Assessment*** se refiere al proceso de recopilar información cuantitativa y cualitativa acerca del aprendizaje estudiantil con distintos propósitos. Se utiliza el término en inglés ya que no hay una traducción adecuada en español (Véase Medina Díaz y Verdejo Carrión, 2006).

6 Utilizo el género femenino o el nombre colectivo para denominar a los maestros, los estudiantes y los administradores escolares. En otras ocasiones utilizaré el género masculino con la intensión de subrayar que las personas del género masculino son la mayoría del colectivo.

7 Los niveles de riego varían dependiendo de las consecuencias atadas a los resultados de las pruebas: de bajo riesgo (***low stakes***), donde no hay consecuencias notables o de alto riesgo (***high stakes***), donde las consecuencias son

severas tales como no otorgar un diploma o cerrar una escuela. También hay que considerar qué o quiénes se afectan (estudiantes, maestros, administradores escolares, escuelas y los sistemas escolares).

8 Utilizo el término ítems para referirme a las preguntas o tareas que incluye una prueba.

9 La ley IDEA (***Individuals with Dissabilities Education Act***) requiere que los estados y distritos incluyan estudiantes de Educación Especial en los programas de pruebas con los acomodos apropiados e informen sus resultados por separado.

10 Comercial se refiere a que la persona, agencia o compañía que construyó o distribuye la prueba tiene un fin de lucro y cobra cierta cantidad de dinero por desarrollarla, administrarla y usarla.

11 Traducción libre de la autora de cada sección de la ley que se cita. Los términos en inglés son para acentuar lo que establece la ley NCLB.

12 Traducción libre del inglés ***limited English proficient students.***

13 Este término ***proficiency*** se ha traducido como proficiencia, por el uso común. Sin embargo, la autora reconoce que no es la traducción adecuada en la lengua española, sino habilidad o pericia.

14 Consultar el artículo de Linn (2003) para un ejemplo de cómo se puede calcular el AYP.

15 Para Puerto Rico se considera como lenguaje materno el español.

16 Expertos se refieren a personas conocedoras y estudiosas del asunto o disciplina en cuestión.

17 Véase la crítica de Medina-Díaz (1998) a esta prueba en el ***Thirteenth Mental Measurements Yearbook***, pp.40-42.

18 No sólo han aumentado los ingresos de las compañías que construyen las pruebas y de las que producen materiales relacionados sino el precio de las viviendas en las zonas cercanas a las escuelas con puntuaciones altas en las pruebas.

Este artículo se recibió en la Redacción de Pedagogía en enero de 2007 y se aceptó para su publicación en mayo del mismo año.