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

Grupo de Trabajo sobre Estándares y Evaluación

¿Por qué las pruebas estandarizadas no miden la calidad educativa?

W. James Popham

Grupo de Análisis para el Desarrollo

¿POR QUÉ LAS PRUEBAS ESTANDARIZADAS NO MIDEN LA CALIDAD EDUCATIVA?

W. James Popham

*Tomado y traducido de **Educational Leadership**, volumen 56, número 6, marzo de 1999, con autorización de ASCD (editores). El autor es profesor emérito de UCLA.*

> ***Los educadores viven una presión casi implacable para demostrar su eficacia. Desafortunadamente, el principal indicador con el cual la mayoría de las comunidades evalúa el éxito del cuerpo docente de una escuela es el desempeño de los estudiantes en pruebas estandarizadas.***

En estos días, si los puntajes que obtiene una escuela en pruebas estandarizadas son altos, la gente piensa que su profesorado es eficaz. Si los puntajes en las pruebas estandarizadas son bajos, se considera que el cuerpo de profesores es ineficaz. En ambos casos, esas evaluaciones pueden ser erradas, porque la calidad educativa está siendo medida con una vara equivocada.

Una de las principales razones por las cuales los puntajes en las pruebas estandarizadas continúan siendo el factor más importante en la evaluación de una escuela es engañadoramente simple. La mayoría de los educadores no comprenden realmente por qué una prueba estandarizada proporciona una evaluación equívoca sobre la eficacia del cuerpo de docentes de una escuela. Y deberían entenderlo.

¿Qué tiene que ver el nombre?

Una prueba estandarizada es cualquier examen que se administra y califica siguiendo un procedimiento estándar predeterminado. Hay dos tipos principales de pruebas estandarizadas: las pruebas de aptitud y las pruebas de logros.

Las pruebas estandarizadas de *aptitud* predicen cuán bien es probable que los estudiantes se desempeñen en algún espacio o nivel educativo subsiguiente. Los ejemplos más comunes son el SAT-I *(Scholastic Aptitude Test)* y el ACT *(American College Testing Assessment)*, los cuales intentan pronosticar cuán bien se desempeñarán los estudiantes secundarios en la universidad.

Pero a la hora de evaluar la eficacia de una escuela, los ciudadanos y los miembros de los consejos escolares se apoyan en los puntajes obtenidos por los alumnos en pruebas estandarizadas de *logros.* En los EEUU, en el ámbito nacional, se usan cinco de esas pruebas: la Prueba de Logros de California, las Pruebas Integrales de Destrezas Básicas, las Pruebas de Destrezas Básicas de Iowa, las Pruebas Metropolitanas de Logros, y las Pruebas de Logros de Stanford.

La misión evaluadora de las pruebas estandarizadas

La gente que diseña las pruebas estandarizadas de logros es extraordinariamente talentosa. Lo que trata de hacer es crear herramientas de evaluación que permitan hacer una inferencia válida sobre los conocimientos y/o las destrezas que posee un estudiante determinado en un área particular de contenidos. Más precisamente, esa inferencia tiene que referirse a normas, de manera que los conocimientos y/o las destrezas relativas de un estudiante puedan ser comparadas con las poseídas por una muestra nacional de estudiantes de la misma edad o grado escolar.

Esas inferencias relativas del status de un estudiante respecto a su dominio de conocimientos y/o destrezas en un área temática o asignatura particular pueden dar mucha información a los padres y educadores. Por ejemplo, piense sobre los padres que descubren que su hijo de 4to. grado se está desempeñando realmente bien en lenguaje (percentil 94) y matemáticas (percentil 89), pero más bien insatisfactoriamente en ciencias (percentil 39) y estudios sociales (percentil 26). Esa información, al echar luz sobre las fortalezas y las debilidades de un niño, puede ser útil no solo para abordar el tema con el profesor, sino

también para brindarle ayuda en casa. De igual manera, si los profesores saben cómo están sus alumnos en comparación con otros del país, pueden usar esta información para planificar una enseñanza apropiada en el salón de clase.

Pero es probable que los chicos en cualquier grado escolar dominen una enorme cantidad de conocimientos y/o destrezas. El tamaño sustancial de la esfera de contenidos que se supone que una prueba estandarizada representa, plantea dificultades genuinas a la labor de los diseñadores de esas pruebas. Si una prueba cubriera en realidad todo los conocimientos y las destrezas de un dominio, sería excesivamente larga.

Por lo tanto, las pruebas estandarizadas de logros a menudo tienen que cumplir su labor de medición con un conjunto mucho menor de ítems que el que se usaría si el tiempo que requiere una prueba no fuera problema. Para salir de esta complicación las pruebas estandarizadas de logros muestrean los conocimientos y/o las destrezas de una esfera o dominio de contenidos. Frecuentemente esas pruebas intentan cumplir su labor evaluadora con solo 40 o 50 ítems -- a veces menos -- en el campo de una asignatura.

La diferenciación precisa como deidad

Los diseñadores de las pruebas estandarizadas de logros tienen la tarea de crear un instrumento de evaluación que, con un puñado de ítems, proporcione interpretaciones válidas, referidas a normas, sobre la situación de un estudiante respecto a una porción sustancial de contenidos. Los ítems que cumplen mejor la labor de diferenciar o discriminar entre los estudiantes son aquellas que aproximadamente la mitad de ellos responden correctamente. Los diseñadores evitan los ítems que son respondidos correctamente por demasiados estudiantes o por muy pocos de ellos.

Haciendo un muestreo cuidadoso de los contenidos y concentrándose en los items que discriminan óptimamente a los estudiantes, los diseñadores de pruebas han producido herramientas de evaluación que son muy buenas para comparar el dominio relativo de contenidos que ha logrado un escolar con los demás escolares del país. Si se asume que el grupo normativo (promedio) nacional es genuinamente representativo del país, los educadores y los padres pueden hacer inferencias útiles sobre los estudiantes.

Una de las más útiles de esas inferencias se refiere a las fortalezas y debilidades relativas de los estudiantes entre distintas asignaturas, como cuando los padres encuentran que su hija brilla en matemáticas pero marcha muy mal en ciencias. También es posible identificar las fortalezas y debilidades relativas de los estudiantes al interior de una misma asignatura si la prueba contiene suficientes ítems para ello. Por ejemplo, si una prueba estandarizada de matemáticas de 45 ítems asigna 15 ítems a cálculo básico, 15 ítems a geometría, y 15 a álgebra, es posible obtener una idea aproximada de las fortalezas y debilidades relativas de un estudiante en esos tres campos de las matemáticas. Sin embargo, es muy frecuente que estas pruebas contengan demasiados pocos ítems como para permitir comparar significativamente fortalezas y debilidades de los estudiantes al interior de una misma asignatura.

Un segundo tipo de inferencia útil que puede basarse en pruebas estandarizadas de logros es aquélla relacionada con el mejoramiento del estudiante a lo largo del tiempo en diferentes asignaturas. Por ejemplo, a un niño se le puede tomar una prueba estandarizada de logros cada tres años. Podríamos ver que los percentiles de desempeño del niño en la mayoría de las asignaturas son similares en cada prueba, pero que en matemáticas parece caer dramáticamente en cada período. Esa es una información útil.

Desafortunadamente, tanto los padres como los educadores suelen atribuir demasiada precisión y exactitud a los puntajes de los estudiantes en las pruebas estandarizadas de logros. Varios factores pueden hacer que los puntajes fluctúen. El que estos puntajes se presenten en cifras (¡a veces incluso con decimales!) no significa que se les puede atribuir una precisión no garantizada. Los puntajes de las pruebas estandarizadas de logros deberían ser tomados como aproximaciones gruesas al status de un estudiante respecto a la esfera de contenidos representada en la prueba.

Para resumir, las pruebas estandarizadas de logros hacen una excelente labor suministrando la evidencia necesaria para hacer interpretaciones referidas a normas respecto a los conocimientos y/o a las destrezas de los estudiantes en relación a aquéllos de los demás estudiantes del país. Esas interpretaciones tienen una utilidad educativa considerable. Las pruebas estandarizadas de logros son, en realidad, bastante notables, si tomamos en cuenta las dimensiones de las esferas o dominios de contenido que están representadas y el número limitado de ítems que los diseñadores de pruebas tienen a su disposición. Hacen lo que se supone deben hacer.

Pero las pruebas estandarizadas de logros no deberían ser usadas para evaluar la calidad de la educación. Eso no es lo que se supone deben hacer.

Midiendo la temperatura con una cuchara

Por varias razones importantes, las pruebas estandarizadas de logros no deberían ser utilizadas para evaluar la calidad de la educación. La razón primordial por la cual los puntajes de los estudiantes en estas pruebas no suministran un indicio preciso de la eficacia de la enseñanza es que cualquier inferencia acerca de la calidad educativa basada en los logros de los estudiantes en las pruebas estandarizadas de logros tiende a no ser válida.

Emplear pruebas estandarizadas de logros para averiguar la calidad educativa es como medir la temperatura con una cuchara. Las cucharas tienen la misión de medir cosas diferentes que el calor o el frío. Las pruebas estandarizadas de logros tienen la misión de medir algo distinto que cuán buena o cuán mala es una escuela. Las pruebas estandarizadas de logros deberían usarse para hacer las interpretaciones comparativas que se supone deben suministrar. No deberían ser usadas para evaluar la calidad educativa. Veamos tres razones significativas por las cuales es totalmente inválido basar inferencias respecto a la calidad de la educación en los puntajes de las pruebas estandarizadas de logros.

Desajustes entre la enseñanza y la medición

Las compañías que diseñan y venden pruebas estandarizadas de logros son todas propiedad de grandes corporaciones. Como todo negocio con fines lucrativos, estas corporaciones intentan generar utilidades para sus accionistas.

Reconociendo la considerable presión por vender pruebas estandarizadas de logros, aquéllos que las comercializan se tropiezan con el difícil dilema que suscita la gran diversidad curricular en los Estados Unidos. Debido a que los distintos estados suelen elegir objetivos educacionales algo distintos (o, para estar a la moda, estándares de contenido distintos), existe la necesidad de elaborar pruebas estandarizadas de logros alineadas apropiadamente con las preferencias curriculares significativamente diferentes de los educadores. El problema se exacerba aun más en los estados en los cuales sus diferentes condados o distritos escolares pueden ejercer decisiones más locales sobre el curriculum.

En un nivel muy general, las metas perseguidas por los educadores en diferentes escenarios son razonablemente similares. Por ejemplo, uno puede estar seguro de que todas las escuelas darán atención al lenguaje, las matemáticas, etc. Pero eso es cierto a un nivel general. En el salón de clase -- el nivel que realmente marca una diferencia para la enseñanza -- hay diferencias significativas en los objetivos educacionales perseguidos. Y eso representa un problema para aquéllos que deben vender las pruebas estandarizadas.

En vista de la sustancial diversidad curricular del país, los diseñadores de pruebas se ven obligados a crear una serie de evaluaciones “talla única”. Pero, como sabemos la mayoría de los que hemos intentado usar ropa talla única, a veces la talla única no le queda bien a todos.

Los diseñadores de estas pruebas hacen lo mejor que pueden al seleccionar los ítems que puedan medir todos los conocimientos y destrezas de un área de contenido que los educadores nacionales consideran importantes. Pero no pueden conseguirlo realmente. Así, pues, las pruebas estandarizadas incluirán siempre muchos ítems que no están alineados con lo que se enfatiza en la enseñanza en un contexto determinado.

Los educadores deben conocer un importante estudio de la Universidad Estatal de Michigan, publicado en 1983 por Freeman y sus colegas, que ilustra el serio desajuste que puede ocurrir entre lo que se enseña en una localidad y lo que se evalúa mediante las pruebas estandarizadas de logros. Estos investigadores seleccionaron cinco pruebas estandarizadas a nivel nacional de logros en matemáticas, y estudiaron sus contenidos para cuarto, quinto y sexto grados. Luego, bajo la hipótesis bastante razonable de que lo que ocurre en la enseñanza en los salones de clase suele estar influenciado por el contenido de los textos que los niños usan, estudiaron cuatro textos ampliamente utilizados en esos grados.

Empleando procedimientos rigurosos de revisión, los investigadores identificaron en las pruebas estandarizadas de logros los ítems que no habían recibido atención significativa en los textos escolares. Concluyeron que los textos escolares no abordaban adecuadamente entre el 50 y el 80% de lo que se medía en las pruebas. Como lo plantearon los investigadores de Michigan, “la proporción de temas presentados en las pruebas estandarizadas que reciben más que un tratamiento superficial en cada texto escolar nunca fue superior al 50%” (p.509).

Bien, si el contenido de las pruebas estandarizadas no es abordado satisfactoriamente en los textos escolares de amplio uso, ¿no es también probable que en un contexto educativo particular las pruebas cubran tópicos que no son abordados en la enseñanza de ese centro? Desafortunadamente, debido a que la mayoría de los educadores no están genuinamente familiarizados con los ingredientes de las pruebas estandarizadas de logros, ellos frecuentemente asumen que si las pruebas estandarizadas sostienen que están evaluando “las capacidades de comprensión de lectura de los niños”, entonces la prueba probablemente concuerde con la manera en que se enseña a leer en esa localidad. Esta supuesta concordancia entre lo que evalúa la prueba y lo que se enseña, resulta bastante a menudo infundada.

Si uno toma tiempo suficiente para leer los materiales descriptivos presentados en los manuales que acompañan las pruebas estandarizadas, encontrará que la descripción de lo que se evalúa suele ser bastante general. Esas descripciones necesitan ser generales para hacer que las pruebas sean aceptables para una nación de educadores cuyas preferencias curriculares son variadas. Pero esas descripciones generales de lo que se evalúa a menudo dan cabida a supuestos alineamientos enseñanza - medición que están lejos de ser ciertas. Y tales desajustes, reconocidos o no, suelen conducir a conclusiones espurias sobre la eficacia de la educación en un lugar determinado, si se utilizan los puntajes de las pruebas estandarizadas como indicador de eficacia educativa. Y esa es la primera razón por la cual no deberían emplearse pruebas estandarizadas de logros para determinar la eficacia de un estado, distrito, escuela, o profesor. Es casi seguro que habrá un desajuste significativo entre lo que se enseña y lo que se mide.

Una tendencia psicométrica a eliminar ítems importantes de las pruebas

Una segunda razón por la cual no deberían usarse pruebas estandarizadas de logros para evaluar la calidad educativa surge directamente del requerimiento de que estas pruebas permiten hacer comparaciones significativas entre estudiantes basadas en sólo un conjunto pequeño de ítems.

Un ítem que es respondido correctamente por cerca de la mitad de estudiantes es el ítem que mejor dispersa los puntajes totales de los estudiantes. Los ítems que son respondidos correctamente por el 40 al 60 por ciento de los estudiantes hacen un buen trabajo en dispersar los puntajes totales de los examinados.

En cambio, los ítems que son respondidos correctamente por un gran número de estudiantes, no contribuyen a dispersar convenientemente los puntajes de las pruebas. Desde la perspectiva de la eficacia de una prueba para proporcionar interpretaciones comparativas, un ítem que es respondido correctamente por el 90 por ciento de los examinados está siendo respondido correctamente por demasiados estudiantes.

Por ello, los ítems respondidos correctamente por el 80 por ciento o más de los examinados suelen no pasar el corte final cuando una prueba estandarizada de logros se diseña por primera vez, y esos ítems probablemente serán desechados cuando se revise la prueba. De ahí que la vasta mayoría de ítems de las pruebas estandarizadas de logros sean ítems de "mediana dificultad".

A consecuencia de la búsqueda de una varianza de puntajes en las pruebas estandarizadas de logros, los ítems en los cuales los estudiantes se desempeñan bien suelen ser excluidos. Sin embargo, estos ítems cubren muy a menudo el contenido al cual, debido a su importancia, los profesores dan mayor énfasis. Así, cuanto más se esfuercen los profesores en enseñar conocimientos y/o destrezas importantes, menos probable será que las pruebas estandarizadas de logros incluyan ítems que los midan. Evaluar la eficacia de la enseñanza de los profesores mediante herramientas de evaluación que deliberadamente evaden los contenidos importantes es fundamentalmente disparatado.

Causalidades confundidas

La tercera razón por la cual no debería usarse pruebas de logros estudiantiles para evaluar la calidad educativa es la más fuerte. Afirmar que puntajes bajos o altos en las pruebas son resultado de la calidad de la enseñanza es ilógico porque el desempeño de los estudiantes en las pruebas estandarizadas de logros está fuertemente influenciado por tres factores, sólo uno de los cuales está vinculado a la calidad educativa.

Para comprender claramente este problema de causalidad confusa, veamos los tipos de ítems que aparecen en las pruebas estandarizadas de logros. Recuérdese que los puntajes se basan en cuán bien responden los estudiantes los ítems de la prueba. Para obtener una idea realmente sólida sobre el contenido de las pruebas estandarizadas, uno necesita examinar los propios ítems.

Los tres ítems ilustrativos aquí presentados son versiones suavemente maquilladas de pruebas estandarizadas de logros actualmente en uso. He modificado ligeramente el contenido de los ítems, sin alterar la esencia de lo que ellos tratan de medir.

El problema de causalidad confusa está relacionado con los tres factores que contribuyen al puntaje de los estudiantes en las pruebas estandarizadas de logros: (1) lo que se enseña en la escuela, (2) la capacidad intelectual innata del estudiante, y (3) el aprendizaje del estudiante fuera de la escuela.

Qué se enseña en la escuela. Algunos de los ítems de las pruebas miden los conocimientos o destrezas que los estudiantes aprenden en la escuela. En ciertas asignaturas tales como las matemáticas, los niños aprenden la mayor parte de lo que saben sobre el tema en la escuela. Pocos padres pasan mucho tiempo enseñando a sus hijos las intrincancias del álgebra o cómo probar un teorema.

Así que si uno mira los ítems en cualquier prueba estandarizada de logros, encontrará un buen número de ítems similares al ítem de matemáticas presentado en la figura 1, que es una versión levemente modificada de un ítem que aparece en una prueba estandarizada de logros para niños del tercer grado.

Isabel tenía 14 peras. Luego regaló 6. ¿Cuál de las siguientes oraciones numéricas podrías utilizar para averiguar cuántas peras le quedaron a Isabel?

A. 14 + 6 = ☐
B. 6 + 14 = ☐
C. ☐ - 6 = 14
D. 14 – 6 = ☐

Figura 1. Item extraído de una prueba de logros estandarizada en matemáticas para tercer grado.

Este ítem de matemáticas ayudaría a los profesores a llegar a una inferencia válida sobre las capacidad de los estudiantes de tercer grado para elegir oraciones numéricas que

coinciden con representaciones verbales de problemas de sustracción. O, este ítem, junto con otros ítems similares relacionados a la adición, multiplicación y división, podría contribuir a una inferencia válida sobre la capacidad del estudiante para elegir varias oraciones numéricas apropiadas a una variedad de problemas de cálculo básico presentados en forma verbal.

Si los ítems en las pruebas estandarizadas de logros midieran sólo lo que realmente se ha enseñado en la escuela, no me opondría al uso de estas pruebas para determinar la calidad educativa. Sin embargo, como se verá pronto, en las pruebas estandarizadas de logros se esconden otros tipos de ítems.

La capacidad intelectual innata de un estudiante. Quisiera creer que todos los niños nacen con idéntica capacidad intelectual, pero no es así. Algunos niños tuvieron más suerte en el momento del reparto de genes. Algunos niños, desde su nacimiento, encontrarán más fáciles las matemáticas que otros. A algunos niños, desde su nacimiento, les serán más fáciles los asuntos verbales que a otros. Si los niños vinieran al mundo habiendo heredado capacidades intelectuales idénticas, los problemas pedagógicos de los profesores serían mucho más simples.

Recientemente, muchos educadores importantes sugieren que hay diversas formas de inteligencia, no sólo una (Gardner 1994). Un niño que nace con menor aptitud para afrontar tareas cuantitativas o verbales, por tanto, puede poseer una mayor inteligencia "interpersonal" o "intrapersonal", pero estas capacidades no son evaluadas por estas pruebas. Los niños difieren en sus capacidades innatas para responder correctamente los tipos de ítems que se encuentran más comúnmente en las pruebas estandarizadas de logros. Y algunos ítems en las pruebas apuntan directamente a medir esa capacidad intelectual.

Considere, por ejemplo, el ítem en la figura 2. Este ítem intenta medir la capacidad del niño para "descifrar" cuál es la respuesta correcta. Pienso que este ítem no mide lo que se enseña en la escuela. El ítem mide lo que traen los estudiantes a la escuela, no lo que allí aprenden.

Si alguien quiere realmente conservar recursos, una buena manera de hacerlo es:

A. Dejar encendidas las luces aun si no se necesita iluminación.
B. Lavar cantidades pequeñas de ropa en la lavadora, en lugar de cantidades mayores
C. Escribir en ambos lados de una hoja de papel.
D. Echar periódicos viejos a la basura.

Figura 2. Un ítem extraído de una prueba estandarizada de logros en estudios sociales para el sexto grado.

Vea cuidadosamente las cuatro opciones de respuesta en el ítem de estudios sociales para alumnos de sexto grado de la figura 2. Lea cada opción y vea si podría ser correcta. Un estudiante "inteligente", sostengo, puede imaginarse que las opciones A, B y D no podrían realmente "conservar recursos" tan bien, de allí que la opción C sea la ganadora. A los chicos más inteligentes les irá mejor con este ítem que a sus compañeros menos inteligentes.

Pero, podría usted estar pensando, ¿por qué incluyen los diseñadores de pruebas estandarizadas esos ítems en sus pruebas? La respuesta es muy simple. Estos tipos de ítem, al tocar destrezas intelectuales innatas que no son fácilmente modificables en la escuela, cumplen una magnífica labor en cuanto a dispersar los puntajes de los examinados. La búsqueda de varianza en los puntajes, junto con las limitaciones de tener que usar pocos ítems en la evaluación escolar, hace que esos ítems se vuelvan atractivos para los diseñadores de pruebas estandarizadas de logros.

Pero los ítems que básicamente miden diferencias en las capacidades intelectuales innatas de los estudiantes obviamente no contribuyen a hacer inferencias válidas acerca "cuán bien se ha enseñado a los niños". ¿Nos gustaría que a todos los niños les fuera bien en esos ítems de "inteligencia innata"? Claro que sí. Pero usar esos ítems para evaluar la eficacia educacional es simplemente erróneo.

Aprendizaje fuera de la escuela. Los ítems más problemáticos en las pruebas estandarizadas miden lo que los estudiantes han aprendido fuera de la escuela. Desafortunadamente, en las pruebas estandarizadas de logros, encontrará más ítems de éstos que los que sospecha. Si los niños proceden de familias aventajadas y de ambientes ricos en estímulos, entonces son más capaces de tener éxito en los ítems de la prueba estandarizada que otros niños cuyos entornos no encajan tan bien con lo que las pruebas miden. El ítem de la figura 3 aclara lo que realmente está siendo evaluado con varios ítems de pruebas estandarizadas de logros.

El fruto de una planta siempre contiene semillas. ¿Cuál de los siguientes no es un fruto?

A. naranja *C. Manzana*
B. calabaza *D. Apio*

Figura 3. Un ítem extraído de una prueba estandarizada de logros en ciencias para el sexto grado.

Este ítem de ciencias de sexto grado le dice a los estudiantes, primero, lo que es un atributo de una fruta (específicamente, que contiene semillas). Luego el estudiante debe identificar lo que "no es una fruta" mediante la selección de la opción sin semillas. Como sabe todo niño que se ha encontrado con un apio, el apio es una planta sin semillas. La respuesta correcta, entonces, para aquéllos que se han topado con los filamentos del apio pero nunca con sus semillas, es claramente la opción D.

Pero, ¿qué hubiera ocurrido si cuando Ud. era un chiquillo, su familia no tenía el dinero para comprar apio en la tienda? ¿Qué hubiera pasado si las circunstancias simplemente no le dieron la oportunidad de tener interacciones significativas con tallos de apio para cuando llegó al sexto grado? ¿Hubiera respondido correctamente el ítem en la figura 3? Y ¿qué bien le hubiera ido si no sabía que las calabazas son esferas que contienen semillas? Si unos niños saben sobre calabazas y apio, es claro que les irá mejor en este ítem que a aquéllos niños que sólo saben sobre manzanas y naranjas. Esa es la manera en que el status socioeconómico de los niños se mezcla con su desempeño en las pruebas estandarizadas de logros. Cuanto más alto status socioeconómico tenga la familia de uno, es más probable que le vaya bien en varios ítems de esa prueba.

Suponga que usted es el director de una escuela en la cual la mayoría de estudiantes proceden de situaciones socioeconómicas realmente bajas. ¿Cómo se desempeñarán, probablemente, sus estudiantes en una prueba estandarizada de logros, si un número importante de ítems de la prueba mide en realidad la riqueza de estímulos de sus ambientes? Correcto, es probable que sus estudiantes no obtengan puntajes muy altos. ¿Significa eso acaso que sus profesores están enseñando deficientemente? Por supuesto que no.

Imaginemos lo contrario, que usted es el director de una escuela rica cuyos estudiantes tienden a tener padres bien educados, de clase alta. Cada año, los puntajes de sus alumnos en las pruebas estandarizadas de logros son deslumbrantemente altos. ¿Significa que su profesorado está haciendo una labor extraordinaria? Por supuesto que no.

Una de las razones principales por las cuales el status socioeconómico de los niños está tan altamente correlacionado con los puntajes de las pruebas estandarizadas es que muchos ítems de las pruebas estandarizadas de logros se concentran en realidad en evaluar conocimientos y/o destrezas aprendidas fuera del colegio -- conocimientos y destrezas probablemente aprendidas más en algunos escenarios socioeconómicos que en otros.

Uno podría preguntarse, nuevamente, ¿por qué es que los diseñadores de pruebas estandarizadas de logros insertan ítems semejantes en sus pruebas? Como siempre, la respuesta es consistente con la misión dominante de medición de esas pruebas, es decir, para dispersar los puntajes de los estudiantes de manera que se puedan hacer interpretaciones finas y precisas referidas a normas. Debido a que existe una variación sustancial en las

situaciones socioeconómicas de los niños, los ítems que reflejan esas variaciones son eficientes para producir variaciones entre estudiantes en los puntajes de las pruebas.

Se acaba de considerar tres factores importantes que pueden influenciar los puntajes de los estudiantes en las pruebas estandarizadas de logros. Uno de esos factores estaba directamente vinculado con la calidad educativa. Pero dos factores no.

¿Qué debe hacer un educador?

He descrito una situación que, desde la perspectiva de un educador, se ve bastante desoladora. ¿Qué puede hacer? Sugiero un triple abordaje al problema. Primero, pienso que se necesita saber más sobre las interioridades de las pruebas estandarizadas de logros. Segundo, pienso que se necesita llevar a cabo una campaña educativa eficaz, de manera que los colegas educadores, los padres de familia, y los diseñadores de políticas educativas comprendan cuáles son realmente las debilidades evaluativas de las pruebas estandarizadas de logros. Por último, pienso que se necesita organizar una forma más apropiada de reunir evidencias basadas en la evaluación.

Aprendiendo sobre las pruebas estandarizadas de logros. Hay demasiados educadores que no han estudiado realmente los ítems de una prueba estandarizada de logros desde los tiempos en que ellos, como estudiantes, estuvieron obligados a responder esos ítems. Pero las inferencias basadas en el desempeño en las pruebas de los estudiantes descansan nada más que en una suma agregada de sus respuestas a los ítems. Los educadores necesitan darse un buen tiempo con las pruebas de logros, analizándolas ítem por ítem para ver lo que realmente están midiendo.

Difundir la noticia. La mayoría de los educadores, y casi todos los padres y miembros de las juntas directivas escolares, piensan que las escuelas deberían ser calificadas en base a los puntajes que obtienen sus estudiantes en las pruebas estandarizadas de logros. Esa gente necesita ser educada. Es responsabilidad de todo educador brindar esa educación.

Si trata de explicar a la opinión pública, a los padres y a los diseñadores de políticas por qué los puntajes de las pruebas probablemente suministran un cuadro engañoso de la calidad educativa, asegúrese de señalar que no está escabulléndose de la responsabilización. Por el contrario, usted debe estar dispuesto a identificar otras evidencias más confiables del logro estudiantil.

Presentar otras evidencias. Si va a argumentar en contra de las pruebas estandarizadas de logros como fuente de evidencia educacional para determinar la calidad de la escuela, y si aún desea ser educacionalmente responsable, entonces necesitará ofrecer de antemano otra forma de evidencia para mostrar al mundo que realmente está haciendo un buen trabajo educacional.

Recomiendo que intente evaluar el grado de dominio por parte de los estudiantes de destrezas cognitivas genuinamente significativas, tales como su habilidad para escribir composiciones efectivas, su habilidad para usar las lecciones de historia para hacer análisis convincentes de problemas actuales y su habilidad para resolver problemas matemáticos de alto nivel.

La reunión de un conjunto de evidencias *ex ante y ex post,* que demuestren el crecimiento sustancial del estudiante en esas ciertas destrezas puede ser verdaderamente persuasiva si las destrezas seleccionadas miden resultados cognitivos realmente importantes, si son consideradas genuinamente significativas por los padres y por los diseñadores de política y si pueden ser abordadas educacionalmente por profesores competentes.

Lo que los profesores necesitan son instrumentos de evaluación que midan destrezas valiosas o conjuntos significativos de conocimientos. Luego los profesores necesitan demostrar al mundo que pueden enseñar a los niños de manera tal que éstos puedan hacer progresos marcados desde la pre instrucción a la post instrucción.

El punto fundamental es éste: si los educadores aceptan la posición de que los puntajes de las pruebas estandarizadas de logros no deberían ser usados para medir la calidad de la enseñanza, entonces deben suministrar otra evidencia confiable que pueda usarse para establecer la calidad de la enseñanza. Una evidencia imparcial, reunida cuidadosamente *ex ante* y *ex post*, sobre la manera en que los profesores promueven conocimientos y destrezas innegablemente importantes podría resolver el problema.

Objetivos correctos, herramientas incorrectas

Los educadores deben ser, definitivamente, responsabilizados. La enseñanza de los hijos de una nación es demasiado importante como para no darle seguimiento. Pero evaluar la calidad educativa mediante el uso de instrumentos de evaluación erróneos es una subversión de la sensatez. Aunque los educadores necesitan producir evidencia válida respecto a su eficacia, las pruebas estandarizadas de logros son herramientas equivocadas para esa tarea.

Referencias

Freeman, D.J., Kuhs, T.M., Porter, A.C., Floden, R.E., Schmidt, W.H., & Schwille, J.R. (1983). "Do textbooks and tests define a natural curriculum in elementary school mathematics?" [Definen los textos y las pruebas un curriculum natural para las matemáticas en la escuela primaria?] en **Elementary School Journal***, 83* (5), 501-513.

Gardner, H. (1994). "Multiple intelligences: The theory in practice" [Inteligencias múltiples: la teoría en práctica] en **Teachers´ College Record***, 95* (4), 574-583.

Nota del autor: Una versión más extensa de este artículo aparecerá en el capítulo final del libro de W. James Popham ***Modern Educational Measurement: Practical guidelines for Educational Leaders*** *[La medición educacional moderna: lineamientos prácticos para líderes educacionales], tercera edición, (próxima aparición); Needham Heights, MA: Allyn & Bacon.*

Puede contactarse con W. James Popham en IOX Assessment Associates, 5301 Beethoven St., Ste.190, Los Angeles, CA 90066 (e-mail:wpopham@ucla.edu).