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

Grupo de Trabajo sobre Estándares y Evaluación

Pruebas de desempeño y reforma educativa

Edward H. Haertel

Grupo de Análisis para el Desarrollo

PRUEBAS DE DESEMPEÑO Y REFORMA EDUCATIVA

Edward H. Haertel

*Tomado y traducido de **Phi Delta Kappan** de mayo de 1999 con autorización de los editores. El autor es profesor de educación en la Universidad de Stanford en California.*

A partir de la publicación de *Una Nación en Riesgo,* una lluvia de informes y declaraciones ha alimentado la percepción popular de que el sistema de educación norteamericano se encuentra en crisis[1]. La disminución a lo largo del tiempo, tanto real como imaginaria, en el desempeño en pruebas como el SAT y la Evaluación Nacional del Progreso Educativo (NAEP); las comparaciones internacionales de la Prueba Internacional del Progreso Educativo y, más recientemente, el Tercer Estudio Internacional en Matemáticas y Ciencias; y las comparaciones con indicadores de logros tales como los de las NAEP establecidos por el Consejo Directivo de Evaluaciones Nacionales han sido publicitados como evidencia de que nuestros problemas educativos no están siendo abatidos. Casi nunca se cuestiona el uso de los puntajes en las pruebas como indicadores de éxito o fracaso educacional. Los puntajes bajos son malas noticias; los puntajes altos son buenas noticias. En la retórica de la reforma educativa, a menudo parece que mejorar el sistema educativo es sinónimo de mejorar los puntajes en las pruebas.

En un clima semejante, la lógica de las pruebas de altas implicancias o costos parece convincente. Evalúen a los alumnos y vean lo que ellos pueden hacer. Responsabilicen a los alumnos o a sus escuelas si no alcanzan el nivel deseado. Antes que dedicarse al micro-manejo de las escuelas, los diseñadores de políticas pueden disponer que se establezcan estándares de contenido y de desempeño para codificar los resultados de aprendizaje esperados y luego dejar que los maestros y los administradores de las escuelas determinen la mejor manera de lograr esos resultados.

Esto da la impresión de ser un plan administrativo racional. Si hay expectativas claras, los maestros sabrán qué se supone que enseñen, los alumnos sabrán qué tanto deben trabajar para alcanzar el nivel deseado y los contribuyentes sabrán si sus escuelas están a la altura de sus expectativas. Si los estándares son apropiados, si los alumnos y los maestros están preparados para aceptar el reto de alcanzarlos, si el período de introducción gradual de la responsabilización es realista, si se encuentran disponibles pruebas confiables y válidas para determinar el alcance del dominio de los alumnos, si los maestros tienen el conocimiento y entrenamiento precisos para ayudar a sus alumnos a enfrentar el reto de los nuevos estándares, si las escuelas no están abrumadas por demandas y exigencias externas, si los materiales y recursos instruccionales necesarios se encuentran disponibles, si se le da la consideración apropiada a los factores externos a la escuela.... entonces un sistema de responsabilización impulsado por la medición debería mostrar con precisión qué alumnos están trabajando y cuáles se están descuidando, qué maestros y escuelas deben ser premiados y cuáles deben ser castigados.

No es difícil comprender por qué la evaluación para la responsabilización es tan del agrado de los diseñadores de políticas. La medición goza de gran apoyo popular[2].

---

[1] David C. Berliner y Bruce J. Biddle, **The Manufactured Crisis** (Reading, Mass: Addison-Wesley, 1995)

[2] Lowell C. Rose y Alec M. Gallup, "The 30th Annual Phi Delta Kappa/Gallup Poll of the Public's Attitudes Toward the Public Schools", **Phi Delta Kappan**, Setiembre 1998; Lowell C. Rose, Alec M.

Demandar más pruebas o pruebas con mayores implicancias o costos es una respuesta visible y dramática a las preocupaciones del público sobre la educación. Mas aún, la idea de que demandar puntajes más altos en las pruebas mejorará la instrucción conlleva la implicancia no muy sutil de que los alumnos, maestros, y administradores simplemente no están haciendo esfuerzos suficientes. Los puntajes se elevarán si se redoblan los esfuerzos. Proponer un nuevo plan de mediciones desvía la atención de los problemas a los que aluden todos esos "si", incluyendo expectativas curriculares contradictorias, preparación docente inadecuada, materiales e infraestructura no apropiados y demografía cambiante de la población estudiantil[3]. Es probable que atacar esos otros problemas tome mucho tiempo y cueste mucho dinero, pero demandar otra nueva prueba no cuesta casi nada. Más aun, una nueva prueba se puede implementar rápidamente, antes de que expiren los períodos de quienes desempeñan un cargo actualmente. Es probable que los puntajes de una nueva prueba sean malos al principio y se eleven en dos o tres años. Como dice Robert Linn, "La imagen demasiado prometedora resultante, coloreada por los beneficios a corto plazo que se ve en la mayoría de los nuevos programas de pruebas, da la impresión de mejoras "justo a tiempo" para las siguientes elecciones"[4].

Si por lo menos no estuvieran presentes todos esos "si". Se ha intentado repetidamente la aproximación a la reforma educativa en base a las pruebas con altas implicancias o costos, generalmente con tristes resultados[5]. Con cada movimiento nuevo de reforma, rebrota la esperanza de que esta vez se evitarán los errores del pasado, que habrán mejoras dramáticas en los resultados del aprendizaje de los alumnos y que los aumentos en los puntajes se generalizarán más allá de las pruebas específicas que se utilizan para responsabilizar a los maestros y alumnos. Un "error del pasado" identificado a lo largo del último decenio ha sido el depender de las pruebas de opción múltiple. Una solución identificada ha sido el confiar en cambiar en las pruebas de desempeño. Las pruebas de desempeño han sido la pieza central en las iniciativas de la reforma educativa estatales y nacionales en los noventa.

A pesar de que existe alguna validez en las acusaciones que se le hacen al uso de las pruebas de opción múltiple y a otros formatos de ítems del tipo de selección de respuestas, yo sostendría que la deficiencia fundamental de las pruebas de altos costos como una herramienta de política educativa yace en la lógica misma de la estrategia

---

Gallup, y Stanley M. Elam, "The 29th Annual Phi Delta Kappa/Gallup Poll of the Public's Attitudes Toward the Public Schools" **Phi Delta Kappan**, Setiembre 1997, pp. 43-44.

[3] Edward H. Haertel, "Student Achievement Tests as Tools of Educational Policy: Practices and Consequences", en Bernard R. Gifford, ed., **Test Policy and Test Performance: Education, Language, and Culture** (Boston: Kluwer Academic Publishers 1989), pp. 35-63.

[4] Ponencia presentada en la reunión anual de la American Educational Research Association, Abril 1998, San Diego, p.2.

[5] Linda Darling-Hammond, "National Standards and Assessments: Will They Improve Education?", **American Journal of Education**, Agosto 1994, pp. 478-510; Gene V. Glass, "Matthew Arnold and Minimal Competence", **Educational Forum**, Enero 1978, pp. 139-44; Daniel M. Koretz, George F. Madaus, Edward H. Haertel, y Albert E. Beaton, **National Educational Standards and Testing: A Response to the Recommendations of the National Council on Education Standards and Testing** (Santa Monica, Calif.: RAND Corporation, 1992); Linn, op.cit.; George F. Madaus, "The Influence of Testing on the Curriculum" en Laurel N. Tanner, ed., **Critical Issues in Curriculum**: 87th NSSE Yearbook (Chicago: National Society for the Study of Education, Universidad de Chicago 1988), pp. 83-121; y Milbrey W. McLaughlin y Lorrie A. Shepard con Jennifer A.O'Day, **Improving Education Through Standards-Based Reform: A Report by the National Academy of Education Panel on Standards-Based Education Reform** (Stanford, Calif.: National Academy of Education, 1995).

reformista, no en la dependencia en un formato de ítem de prueba versus otro. Mientras los educadores y los diseñadores de políticas empiezan a ver que las pruebas de desempeño con altas implicancias o costos pueden haber sido excesivamente publicitadas como una herramienta de la reforma educativa, hay dos lecciones por aprender.

Primero, no es posible encontrar una solución en el retorno a los errores del pasado -- simplemente reemplazando las pruebas de desempeño por pruebas de opción múltiple -- o en una movida hacía un nuevo tipo de ítem. En lugar de esto debe examinarse y reconsiderarse el argumento de las pruebas para la responsabilización.

Segundo, el fracaso de las pruebas de desempeño con altas implicancias o costos para efectuar reformas educativas amplias y profundas no debería desvirtuar el valor real de las pruebas de desempeño utilizadas con fines *instruccionales*. Se puede obtener los beneficios esperados de evaluar a los alumnos utilizando tareas complejas, integradoras, y activas que borran la distinción entre la evaluación y la enseñanza cuando los mismos maestros hacen uso hábil de buenas pruebas de desempeño en sus propios salones o aulas.

La lógica de la reforma impulsada por la medición

A menudo, las razones fundamentales con las que se argumenta a favor de la reforma impulsada por la medición empiezan con una crítica de los anteriores programas de pruebas con altas implicancias o costos[6]. Se dice que las pruebas de opción múltiple se empezaron a usar ampliamente, en gran parte, porque no era caro elaborarlas, administrarlas y calificarlas. La información que proporcionaban cumplía los criterios psicométricos de confiabilidad y validez, prediciendo, por ejemplo, las calificaciones en la universidad. Una vez que se daba cuenta de los efectos de su baja confiabilidad, generalmente se encontraban altas correlaciones entre las pruebas de opción múltiple y exámenes menos objetivos. Estas correlaciones significaban que, si bien las pruebas de opción múltiple y las alternativas más costosas realmente no medían lo mismo, ellas clasificaban a los examinados de manera muy semejante, de modo que las pruebas de opción múltiple proporcionaban esencialmente la misma información que se podía obtener utilizando procedimientos más costosos y prolongados.

Se dice que todo hubiera salido bien si las pruebas de selección de respuestas hubiesen servido realmente sólo como indicadores del desempeño educacional. Pero cuando se añadieron premios y sanciones a esas pruebas, obtener puntajes altos se convirtió en un fin en sí mismo, a lo que inevitablemente siguieron distorsiones en la enseñanza en clase. Cuando los periódicos clasificaron a las escuelas según sus puntajes promedio en las pruebas, los administradores de las escuelas instaron a los maestros a dedicar más tiempo y esfuerzos a la preparación para las pruebas. La mejor manera de obtener puntajes altos en las pruebas de opción múltiple era tocar cuanto trozo aislado de información que fuese posible, con la esperanza de cubrir las que aparecerían en el examen. De este modo, las pruebas externas con altas implicancias o costos condujeron la enseñanza en aula hacia el uso de hojas de trabajo consistentes en hojas y más hojas de ejercicios para rellenar espacios en blanco y elegir la respuesta correcta.

[6] Joan L. Herman, Large-Scale Assessment in Support of School Reform: Lessons in the Search for Alternative Measures (Los Angeles: Center for the Study of Evaluation, Universidad de California, 1997).

Inevitablemente, la evaluación en aula vino a parecerse a las pruebas externas, en detrimento de objetivos educacionales más importantes tales como el razonamiento de mayor nivel de complejidad, la solución de problemas del mundo real y otros.

El corolario de esta historia está fielmente representado en el aforismo "Lo que evalúas es lo que obtienes". Si las pruebas de opción múltiple pueden guiar la enseñanza en la dirección equivocada, ¿por qué no utilizar, entonces, un mejor tipo de pruebas para empujar la enseñanza en la dirección correcta? Se pidieron pruebas para las cuales sería *deseable* que los maestros entrenaran a los alumnos a rendirlas[7]. Si las pruebas de altas implicancias o costos fuesen auténticas, si comprometieran a los alumnos en la solución de problemas del mundo real y demandaran un razonamiento de mayor nivel de complejidad que lo que pueden medir los ítems de opción múltiple, entonces la enseñanza se alinearía pronto con metas educacionales más valiosas.

Los diseñadores de políticas no fueron los únicos en abrazar las pruebas de desempeño como herramienta para la reforma educativa. Los promotores de diversas reformas curriculares vieron las pruebas de desempeño como un medio para comunicar y difundir nuevas visiones sobre el aprendizaje escolar. Estas formas nuevas de evaluación fomentarían un involucramiento activo tanto en aprender como en demostrar lo que se ha aprendido. Servirían como modelos de actividades instruccionales sólidas. Las expectativas se elevarían a medida que los maestros vieran evidencias concretas de que los alumnos lograran resolver más problemas complejos que lo que ellos hubieran imaginado posible, incluso problemas que exigieran una labor de varios días para resolverlos. Los padres llegarían a valorar los tipos de aprendizaje que se producirían cuando grupos de niños trabajaran juntos para diseñar un experimento o investigar algún tema que realmente les interesase. El tiempo de clase sería mejor invertido, a medida que desapareciera la línea divisoria entre la enseñanza y la medición de aprendizajes. Los alumnos mostrarían mayor motivación y los maestros adoptarían espontáneamente estrategias instruccionales más ilustradas.

Se esperaba que con las pruebas de desempeño los puntajes promedio serían más iguales entre alumnos de comunidades ricas y de comunidades pobres y entre alumnos no-hispanos blancos y los alumnos hispanos y negros. A pesar de décadas de investigación sobre los sesgos de los ítems y los sesgos de las pruebas, persiste la sospecha de que el contenido o el formato de las pruebas de opción múltiple favorece injustamente a los estudiantes de aquellos grupos que tradicionalmente obtienen mayores puntajes. Se esperaba que los grupos que tradicionalmente obtenían puntajes más bajos se desempeñarían tan bien como cualquiera de los otros si se les daba la oportunidad de demostrar directamente lo que realmente sabían y podían hacer.

Nótese que de acuerdo con esta argumentación, la evaluación externa guía y la práctica en aula obedece. En realidad, la historia es más complicada. Las pruebas de altas implicancias impuestas desde fuera ejercen cierta influencia en el currículum y la

---

[7] Lauren B. Resnick y Daniel P. Resnick, "Assessing the Thinking Curriculum: New Tools for Educational Reform", en Bernard R. Gifford y Mary C. O'Connor, eds., **Changing Assessments: Alternative Views of Aptitude, Achievement, and Instruction** (Boston: Kluwer Academic Publishers, 1992), pp.37-75; y Grant Wiggins, "Creating Tests Worth Taking", **Educational Leadership**, Mayo 1992, pp. 26-33.

enseñanza, pero su poder en ese aspecto ha sido exagerado[8]. Existen explicaciones alternativas sobre el uso en clase de hojas de trabajo, la enseñanza de destrezas fuera de contexto y una sobre-enfatización en la memorización de datos a expensas de la resolución y la aplicación de problemas más complejos. Las pruebas de opción múltiple de altas implicancias desempeñaron un papel, pero existieron justificaciones independientes para introducir destrezas componentes de manera aislada y para diferir el razonamiento más complejo hasta que se dominaran las destrezas básicas.

Según explican Lauren y Daniel Resnick, las pruebas de opción múltiple concuerdan bien con la psicología educativa conductista dominante hasta principios de los setenta[9]. Dos supuestos básicos de gran parte de la psicología educacional de ese entonces eran la "desagregabilidad" o "descomponibilidad" y la "descontextualización". Los diseñadores de currícula que se basaron en principios conductistas, buscaban construir desempeños complejos a partir de componentes más básicos. Se creía que la clave para ayudar a todos los niños a dominar las complejidades de la lectura, escritura, matemáticas y otras materias escolares, era enseñar (y evaluar) cada destreza componente en la secuencia apropiada. Las destrezas aprendidas de manera aislada (fuera de contexto) podrían luego ser entrelazadas y aplicadas para llevar a cabo tareas significativas. Los objetivos de aprendizaje se definían en términos de conductas observables y se debía demostrar el dominio de los pre-requisitos antes de introducir destrezas más complejas. Las preguntas de opción múltiple y de "relleno de espacios" en blanco parecían funcionar bien para evaluar destrezas componentes al tiempo que se las iba enseñando. Los objetivos del aprendizaje se escribían en términos de conductas observables y a menudo se leían como especificaciones para las pruebas.

Las razones por las cuales en las aulas norteamericanas se enraizaron tanto las hojas de trabajo, los formatos de ítems de pruebas objetivas, etc., tuvieron menos que ver con las pruebas externas de altas implicancias o costos que con el diseño de materiales educativos y con la formación inicial y la capacitación de los maestros. Los investigadores educacionales y los especialistas en currículo fomentaron conscientemente estas prácticas. Se educó a una generación de maestros en la teoría y métodos de la evaluación referida a criterios. Cambiar la currícula y la enseñanza ha requerido siempre más que sólo introducir alguna nueva variedad de pruebas de altas implicancias.

El conductismo se opacó y las pruebas referidas a criterios se volvieron menos populares como estrategia de administración de la enseñanza. Hoy en día se enfatiza menos la enseñanza y evaluación de destrezas componentes aisladas y se da mayor reconocimiento a que las destrezas aprendidas en un contexto pueden ser inaccesibles en otro. Si bien si los adultos pueden ver fácilmente de qué manera el ejercicio y la práctica matutinos deberían apoyar la lectura o la resolución de problemas en la tarde, los estudiantes jóvenes tienen una capacidad asombrosa para no hacer esas conexiones. En efecto, durante el muy publicitado declive de los puntajes en los años setenta y principios de los ochenta, en realidad los puntajes de las pruebas de opción múltiple en

[8] Lorrie A. Shepard, Insights Gained from a Classroom-Based Assessment Project (Los Angeles: Center for the Study of Evaluation, Universidad de California, 1997); Mary L. Smith, Reforming Schools by Reforming Assessment: Consequences of the Arizona Student Assessment Program (ASAP): Equity and Teacher Capacity Building (Los Angeles: Center for the Study of Evaluation, Universidad de California, 1997).

[9] Resnick y Resnick, pp.41-44.

los primeros grados de educación básica se estaban elevando, mientras que el desempeño de los alumnos mayores en tareas más complejas estaba decayendo.

Cuando cundió la desilusión con la idea de que las destrezas componentes, una vez adquiridas, de alguna manera se integrarían espontáneamente en desempeños proficientes, se dio un cambio que se apartaba del modelo implícito de las pruebas como indicadores de destrezas componentes hacia un modelo de pruebas como muestras directas de desempeños de criterios deseados. Las pruebas de desempeño dejaron de ser vistas como mejores instrumentos para medir los mismos constructos e incluso como instrumentos para medir constructos más importantes. Más bien, fueron concebidas como demostraciones o logros a ser valorados por sí mismos[10].

Ese cambio trae consecuencias importantes para la lógica de hacer inferencias sobre dominios más amplios de tareas partiendo de los puntajes de las pruebas[11]. Si los ítems de las pruebas son diseñados para ser indicadores de dimensiones subyacentes de la proficiencia adquirida (constructos), entonces, *ceteribus paribus*, los puntajes de las pruebas deberían predecir el desempeño en cualquiera de las demás tareas que dependen de los mismos constructos. Si, en cambio se eligen tareas de evaluación por su valor intrínseco, es más difícil decir exactamente qué es lo que los puntajes deben predecir. No existe una justificación clara para esperar que esos puntajes predigan algo que no sea el desempeño en tareas similares, de algún modo, a la prueba misma.

Con respecto a esto, Samuel Messick ha delineado una distinción útil refiriéndose a pruebas de desempeño impulsadas por constructos versus pruebas de desempeño impulsadas por tareas. La diferencia es de grado; siempre existe cierto interés en medir constructos. Aun así, el cambio en el enfoque de medir constructos hacia el de demostrar desempeños valiosos tiene consecuencias significativas para la evaluación del desempeño a gran escala. Messick observa que "*la aproximación a la evaluación del desempeño centrada en la tarea se encuentra en peligro de adaptar los criterios para la calificación y las rúbricas a las propiedades de la tarea y de representar cualquier constructo reducido en formas dependientes de la tarea que pueden limitar la generalizabilidad"*[12]. En realidad, la generalizabilidad limitada de las pruebas de desempeño está bien documentada[13].

El cambio desde pruebas de selección de respuestas impulsadas por constructos hacia pruebas de desempeño impulsadas por tareas se aprecia con mayor claridad en los sistemas de evaluación basados en portafolios. Con las pruebas impulsadas por

---

[10] Wiggins, op.cit.

[11] Edward H. Haertel, "Construct Validity and Criterion-Referenced Testing", **Review of Educational Research**, Vol.55, 1985, pp. 23-46.

[12] Samuel Messick, "Validity of Performance Assessments", en Gary W. Phillips, ed., **Technical Issues in Large-Scale Performance Assessment** (Washington, D.C.: National Center for Education Statistics, 1996), pp. 1-18.

[13] Eva L. Baker, Harold F. O'Neil, y Robert L. Linn, "Policy and Validity Prospects for Performance-Based Assessment", **American Psychologist**, Diciembre 1993, pp. 1210-18; Stephen B. Dunbar, Daniel M. Koretz, y H.D. Hoover, "Quality Control in the Development and Use of Performance Assessments", **Applied Measurement in Education**, Vol. 4, 1991, pp. 289-303; Robert L. Linn, "Educational Assessment: Expanded Expectations and Challenges", **Educational Evaluation and Policy Analysis**, Vol. 15, 1993, pp. 1-16; Linn, "Assessment and Accountability"; y Richard J. Shavelson, Gail P. Baxter, y Jerry Pine, "Performance Assessments: Political Rhetoric and Measurement Reality", **Educational Researcher**, Mayo 1992, pp. 22-27.

constructos se asume que los ítems proporcionados son una muestra de un dominio o campo más amplio de ítems que podrían haberse escogido en su lugar. Este supuesto garantiza la inferencia estadística que se realiza sobre el desempeño predecible en todo campo hipotético de items que representa una prueba partiendo del desempeño de un examinado en la prueba. Sin embargo, con los portafolios los estudiantes participan seleccionando las muestras de su trabajo a ser presentado. Cada componente del portafolio puede ser una buena demostración de proficiencia, pero cualquier inferencia estadística al campo más amplio de potenciales componentes de ese portafolio es problemática.

En resumen, las pruebas de opción múltiple y otras pruebas objetivas de selección de respuestas son claramente inadecuadas como indicadores únicos de los resultados esperados de la enseñanza escolar. Las pruebas de desempeño pueden medir ciertos tipos importantes de resultados que las pruebas de opción múltiple no pueden medir, pero también han provocado un cambio en la perspectiva de la medición que podría limitar la generalizabilidad.

Las pruebas externas de desempeño versus las de desempeño en el aula

Aun si las pruebas de responsabilización tuvieran mayor poder para modelar el currículo y la enseñanza, e incluso si la generalizabilidad de las pruebas de desempeño no fuera importante, quedarían en pie serias objeciones a la estrategia de reforma con pruebas de desempeño de altas implicancias o costos. Cualquier prueba percibida como de "altas implicancias" tiene el potencial de reducir la currícula y de dirigir la enseñanza en aula hacia la demostración del conocimiento en las formas particulares que demanda la prueba. El significado de los puntajes de las pruebas puede cambiar y la validez puede mermar a medida que los maestros enseñen para mejorar el rendimiento, aun si ello fuera para la mejor de las pruebas. De este modo se coloca un peso enorme en los supuestos de que las pruebas externas de desempeño representan realmente indicadores globales, válidos y robustos de los resultados que se desean de la enseñanza[14]. Pero existen razones serias para preguntarnos si las pruebas externas de desempeño pueden cumplir esos supuestos.

Es fácil describir el formato de una prueba de opción múltiple. Las preguntas de opción múltiple se parecen entre si, ya sea si aparecen en una prueba impuesta por un estado o un distrito, o si aparecen en una hoja de trabajo creada por un maestro. Es probable que con la prueba externa de opción múltiple las propiedades psicométricas de los ítems estén mejor documentadas; se pueden indicar las respuestas sombreando los círculos en una hoja de respuestas y puede haber mayor formalidad al leer las instrucciones, al responder las preguntas de los alumnos y al asignar un período de tiempo a la administración. Aun así, el ítem de opción múltiple es parecido en cualquiera de los dos contextos.

Es más difícil describir una prueba de desempeño. El término abarca muchos tipos diferentes de tareas y procedimientos para la calificación[15]. Particularmente, las tareas de desempeño en las pruebas externas, a diferencia de las pruebas en clase, son a menudo radicalmente diferentes -- tanto, que las pruebas externas de desempeño con

[14] Madaus, op. cit.
[15] Edward H. Haertel y Robert L. Linn, "Comparability", en Phillips, pp. 59-78.

altas implicancias pueden servir como modelos inadecuados para la enseñanza en clase y pueden de hecho socavar la comunicación sobre las metas de la reforma curricular[16]. Las tareas de las pruebas externas pueden no comprometer el interés de los niños y no llegar a mostrar lo que los niños saben y pueden hacer en realidad, particularmente los niños de grupos tradicionalmente mal atendidos. Las tareas pueden quedar cortas en lo que se refiere a ejemplificar pruebas para las cuales quisiéramos que enseñaran los maestros.

Las diferencias más obvias entre las evaluaciones del desempeño en el aula y las pruebas externas de desempeño provienen de las restricciones prácticas que se imponen en los programas de pruebas externas de altas implicancias. Aunque algunas pruebas externas, especialmente en escritura y lenguaje, pueden ser administradas a lo largo de varios días, la mayoría se limita a un breve período de evaluación. Cualquier material que sea necesario, salvo los materiales y artículos básicos disponibles en toda escuela, debe ser traído al lugar de la evaluación. El registro calificable se limita en gran medida a las respuestas que los estudiantes pueden escribir en un corto período de tiempo, debido a que la calificación se hace en algún lugar centralizado. Es probable que no sean factibles los discursos, las actuaciones, los proyectos y las composiciones largas. Las pesadas demandas de redacción resultantes pueden ser especialmente problemáticas para los estudiantes a quienes les falta proficiencia en inglés y para aquéllos a quienes les falta motivación.

Ha habido algunos intentos por salvar estas limitaciones, como por ejemplo en Vermont y Kentucky, haciendo que los alumnos armen portafolios de textos escritos por ellos o de soluciones a problemas de matemáticas, que luego son reunidos y evaluados. Generalmente, los maestros reportan efectos positivos modestos en sus prácticas de enseñanza como resultado de estas pruebas, pero los puntajes de los portafolios han mostrado cualidades psicométricas pobres[17]. Más aun, se ha cuestionado si los portafolios de altas implicancias realmente representan el trabajo de los propios alumnos[18].

Existen diferencias más significativas aun en los contextos de las pruebas externas y los de las del aula, más allá de cuestiones de formato, tiempo y calificación. La prueba de desempeño que un maestro crea o selecciona se basa en la instrucción precedente. Ésta puede requerir cierta aplicación de destrezas o ideas en contextos nuevos, pero está referida de manera apropiada al material que los alumnos han tenido oportunidad de aprender. Las pruebas de desempeño externo están referidas apropiadamente no a la enseñanza que se ha impartido sino al marco curricular que define las metas propuestas para la enseñanza. Una prueba en aula que incluye material no enseñado sería considerada injusta incluso si el material estuviese especificado en algún marco curricular. Una prueba externa que omitió material del marco debido a que no fue enseñado sería considerada no válida para casi cualquier propósito.

---

[16] Smith, op. cit.

[17] Daniel Koretz et al., "The Vermont Portfolio Assessment Program: Findings and Implications", **Educational Measurement: Issues and Practice**, Otoño 1994, pp. 5-16; Smith, op. cit.; y Brian M. Stecher et al., **The Effects of Standards-Based Assessment on Classroom Practices: Results of the 1996-97 RAND Survey of Kentucky Teachers of Mathematics and Writing** (Los Angeles: Center for the Study of Evaluation, Universidad de California, 1998).

[18] Herman, p. 25.

Ya que la evaluación en aula puede prepararse en base a la enseñanza precedente, puede inspirarse en habilidades componentes, el vocabulario especial, las rúbricas de calificación y conocimientos de contenidos que han construido el maestro y los alumnos a través del tiempo. Cualquier demostración de razonamiento (o "el razonamiento de mayor nivel de complejidad") depende de cierta base previa de conocimientos; el razonamiento debe ser *acerca* de algo. Si una clase de ciencias ha estudiado la ecología de un estanque y otra ha estudiado la ecología de una pradera, los alumnos de cualquiera de las dos aulas podrían ser capaces de elaborar una cadena alimenticia o de razonar sobre las consecuencias de introducir ciertas especies exóticas, pero las preguntas necesarias para captar esas comprensiones serían distintas. Si dos clases de inglés estudiaron novelas diferentes, los alumnos de cualquiera de las clases podrían ser capaces de demostrar su habilidad para analizar el desarrollo de los personajes o de reconocer la ironía, pero, nuevamente, las preguntas particulares necesarias para captar sus comprensiones serían distintas.

El maestro de aula puede utilizar términos y conceptos presentados anteriormente cuando formula preguntas. Además, si los alumnos han tenido la oportunidad de practicar respondiendo a preguntas similares, es más probable que comprendan qué es lo que se espera a modo de respuesta. Una prueba externa de desempeño puede incluir un guión que describa un ecosistema o un pasaje corto para que los alumnos lo lean y analicen, pero estos estímulos probablemente resultarían más pobres que el material que los alumnos tienen de su propio trabajo en clase. Al formular preguntas en una prueba externa y al explicar a los alumnos cómo deben responder, es poco lo que se puede dar por sentado a modo de antecedentes comunes. Como resultado, las pruebas externas de desempeño se pueden inclinar hacia la medición de "viveza", experiencia con las pruebas o rapidez para la lectura, en lugar de inclinarse hacia la medición del razonamiento y la reflexión.

Los puntajes de las pruebas externas deben ser comparables entre escuelas y a través del tiempo. Por lo tanto, los puntajes asignados no deben depender de la identidad de quien califica ni deben estar influenciados por factores contextuales[19]. Como consecuencia, las rúbricas de calificación para las pruebas externas tienden a focalizarse en aspectos específicos y bien definidos de las respuestas a tareas particulares de las pruebas. Dichas rúbricas ayudan a garantizar la objetividad pero pueden no captar la esencia del desempeño[20]. Las rúbricas específicas a las tareas hacen difícil que alumnos y maestros logren generalizaciones acertadas sobre los criterios de excelencia en la siguiente tarea con la que se encuentren. Por el contrario, los puntajes de las evaluaciones de aula sólo necesitan reflejar los juicios de los maestros sobre la calidad del trabajo de un salón, en una única ocasión. El desempeño de los alumnos puede ser evaluado en relación a criterios más globales. Por razones de equidad, el trabajo de estudiantes diferentes debe ser evaluado de acuerdo a los mismos estándares, pero el maestro es un observador privilegiado, capaz de interpretar y responder al trabajo de los alumnos en el contexto de información colateral[21].

---

[19] Haertel y Linn, op.cit.

[20] Messick, op.cit; y W. James Popham, "What's Wrong -- and What's Right -- with Rubrics", **Educational Leadership**, Octubre 1997, pp.72-75.

[21] Pamela A. Moss et al., "Portfolios, Accountability, and an Interpretive Approach to Validity", **Educational Measurement: Issues and Practice**, Otoño 1992, pp. 12-21.

Por último, el conflicto entre la responsabilización y las metas de la enseñanza interfiere en el valor que pueden tener las pruebas externas para el mejoramiento de la enseñanza en el aula[22]. En el aula, el maestro puede ofrecer asistencia, idealmente proporcionando a los alumnos múltiples oportunidades para dominar las tareas asignadas. Esa función está reñida con el rol de administrador de pruebas. Es común que las pruebas de desempeño en el aula sean anunciadas con anticipación, con tiempo suficiente para estudiar y prepararse. La prueba de desempeño externa posiblemente se guarde en secreto, a menudo tomando por sorpresa tanto a los alumnos como al profesor. Al concluir una prueba de desempeño en el aula, los alumnos frecuentemente presentan sus trabajos ante una audiencia conformada por sus pares, sus maestros e, incluso, sus padres. Son motivados a demostrar lo que pueden hacer y pueden contar con recibir puntajes y retroalimentación individuales. Con las pruebas externas la audiencia es anónima y, en la mayoría de casos, los alumnos no pueden contar con recibir ninguna información sobre su desempeño individual.

En conjunto, las diferencias de formato, contexto, relación con la enseñanza precedente, calificación y objetivos entre las pruebas de desempeño en el aula y las externas son tan grandes como para hacer a estas últimas modelos poco prometedores para la enseñanza o las actividades evaluativas en el aula.

Conclusiones

Las estrategias para el mejoramiento escolar que se apoyan en las pruebas de altas implicancias o costos se basan en una serie de supuestos que carecen de fundamento. Las pruebas de opción múltiple tienen serias limitaciones, pero esas limitaciones no explican el fracaso de los esfuerzos pasados de reformas impulsadas por la medición.

Se ha hecho una propaganda excesiva sobre el poder que tienen las pruebas de altas implicancias de influir en la enseñanza en el aula. Las pruebas de selección de respuestas se volvieron comunes en las aulas porque eran parte de una red compleja de creencias y prácticas; su status preferencial no fue sólo una reacción a las implicancias atadas a las pruebas externas de opción múltiple. Cambiar los formatos de las pruebas de altas implicancias producirá algunos cambios, pero éstos serán limitados.

El cambio de las pruebas de opción múltiple a las pruebas de desempeño, especialmente evaluaciones basadas en portafolios, ha traído consigo un cambio sutil pero importante en la filosofía de la medición. La selección de las tareas de la evaluación sobre la base de su valor intrínseco complica las inferencias de los puntajes de las pruebas referidas a otras situaciones, especialmente si los alumnos y los maestros participan en dicha selección, como es el caso de los portafolios. Las garantías tradicionales que se requieren para interpretar y generalizar a partir de los puntajes en las pruebas se debilitan a medida que las interpretaciones de los constructos se tornan menos relevantes y las tareas evaluativas llegan a ser consideradas como demostraciones apreciables por sí mismas. La generalizabilidad limitada de las pruebas de desempeño está bien documentada. Tanto las generalizaciones sustantivas sobre el significado de puntajes altos como las generalizaciones estadísticas sobre otras tareas se ven amenazadas si las rúbricas de calificación se vuelven demasiado específicas a las tareas.

[22] Smith, op. cit.

Por último, el mismo término “prueba de desempeño” es elusivo. Los requisitos intrínsecos para programas de pruebas de altas implicancias han llevado a la creación de una especie de tarea de desempeño que guarda poca semejanza con las pruebas de desempeño de alta calidad que los maestros son capaces de utilizar en su enseñanza diaria. Así, haciendo a un lado los asuntos psicométricos, se debe cuestionar si las pruebas externas de desempeño pueden servir siquiera como modelos de actividades instruccionales valiosas.

Hay mucho que celebrar en el cambio de la dependencia casi exclusiva de las pruebas de selección de respuestas hacia una combinación de métodos, incluyendo las pruebas de desempeño. No hay garantía alguna, sin embargo, de que los beneficios potenciales de las pruebas de desempeño se vayan a materializar. Al margen del valor de las pruebas de desempeño en aula, una estrategia para la reforma impulsada por la medición que se apoye en las pruebas de desempeño para orientar el currículo y la enseñanza parece estar condenada al fracaso.