INEEd
Instituto Nacional de Evaluación Educativa

Plan estratégico 2013-2016

Anexo III
Evaluación de individuos y evaluación del sistema educativo

*Versión aprobada por Comisión directiva el 28 de febrero de 2013.

**En las publicaciones del INEEd se utiliza el masculino genérico por un criterio de economía de lenguaje y para que su lectura sea más fluida, sin ninguna connotación discriminatoria.

Un principio fundamental a tener en cuenta es que el diseño de una evaluación depende del propósito que se persiga. En particular es fundamental distinguir las evaluaciones que pretenden llegar a juicios sobre individuos (sean estos alumnos, docentes, centros escolares u otros) de las que buscan llegar a conclusiones sobre un sistema educativo en conjunto, como tal.

Actualmente en algunos países hay problemas por la confusión del uso que se puede hacer correctamente de evaluaciones del aprendizaje que tienen como propósito dar información a cada escuela y cada alumno, así como a sus maestros y sus padres, para ajustar sus respectivos esfuerzos para mejorar el aprendizaje ulterior, en contraste con las evaluaciones cuyo propósito es informar sobre la situación promedio del sistema educativo, para sustentar decisiones de política. En el primer caso es necesario aplicar las pruebas a todos los alumnos (censo), mientras que en el segundo basta con hacerlo a una muestra representativa, lo que permite usar instrumentos que midan aspectos más complejos.

Valorar integralmente el desempeño de un solo alumno implica considerar todas las áreas del currículo; los aspectos cognitivos y no cognitivos, simples y complejos; el rendimiento al inicio, a lo largo y al final del ciclo; los factores que favorecen u obstaculizan el avance, etc. Una valoración así solamente puede ser bien hecha por un profesional que tenga contacto con el alumno a lo largo del ciclo escolar: el maestro. La evaluación en aula, sin embargo, aunque es la más completa, tiene una limitación derivada de su misma naturaleza integral y contextualizada: no puede agregarse para ofrecer una visión de la situación educativa a gran escala.

Las pruebas estandarizadas permiten comparar confiablemente grandes grupos, pero no pueden cubrir todos los aspectos que puede atender la evaluación hecha en el aula por los maestros. Por ello las pruebas pueden complementar la evaluación del maestro, pero no sustituirla. Las limitaciones de una prueba censal son mayores, obviamente, que las de una que se aplica solo a una muestra. En particular, en la práctica es imposible un control adecuado de una aplicación censal, mientras que es posible conseguirlo con una muestra.

De manera análoga, la valoración integral de la calidad de un maestro solamente podrá resultar del contacto amplio de profesionales competentes con el evaluando, para reunir evidencias sólidas de su desempeño, mediante entrevistas, observaciones, portafolios de evidencias y otros instrumentos. Las pruebas estandarizadas del rendimiento de los alumnos pueden aportar elementos para la evaluación de sus maestros, en particular utilizando los modelos llamados de valor agregado, pero en la práctica es difícil utilizarlos y, en todo caso, sus resultados solos no son suficientes, por los numerosos factores que inciden en el rendimiento escolar.

Igualmente, valorar de manera integral la calidad de una escuela solo puede ser el resultado de un amplio contacto con el plantel por parte de personas capaces de observar, registrar, sistematizar y valorar evidencias de las múltiples facetas de la calidad educativa en ese nivel. Los sistemas educativos suelen contar para ello con la figura del supervisor o inspector. Reducir la evaluación de la calidad de las escuelas a ordenamientos basados en los resultados de sus alumnos en pruebas en gran escala,

sin tener en cuenta otros elementos y sin asegurar la confiabilidad y validez de los resultados, es muy inapropiado y llevará a distorsiones perniciosas para el funcionamiento de las escuelas.

Los comentarios anteriores se aplican a evaluaciones que servirán para tomar decisiones y emprender acciones que afectarán a sujetos individuales, sean alumnos, maestros o escuelas. Pero si se quiere evaluar el sistema educativo como un todo, entonces tienen sentido acercamientos a alumnos, maestros o escuelas que utilicen instrumentos que se apliquen a muestras representativas de unos u otras. Esos acercamientos no ofrecen bases sólidas para tomar decisiones sobre individuos (como aprobar o reprobar a un alumno, o conceder un estímulo a un maestro o una escuela), pero sí elementos valiosos para valorar la calidad del sistema como tal en relación con esos componentes.

Un elemento más es indispensable para que la evaluación educativa sirva para mejorar la calidad: que los resultados se difundan para que todos los actores interesados en la educación los conozcan y puedan participar en las decisiones y en las acciones de mejora que se emprendan en consecuencia.

Pero no cualquier forma de difusión de resultados es adecuada, y una que no lo sea puede tener un impacto negativo sobre escuelas, maestros y alumnos. El abuso de los ordenamientos de escuelas con base en los resultados obtenidos por sus alumnos en pruebas estandarizadas (los llamados rankings de escuelas) es un ejemplo de que este riesgo no es imaginario, sino tan real que está presente ya en sistemas educativos importantes, como los de Estados Unidos y México. Esos ordenamientos de escuelas son una forma atractiva de presentar resultados de pruebas estandarizadas, pero la idea de que ofrecen un orden inequívoco de la calidad de las escuelas consideradas carece de sustento sólido.

Las pruebas estandarizadas, en especial si se aplican censalmente, tienen niveles de precisión gruesos y solamente consideran el rendimiento en unas áreas y temas del currículo que se pueden medir con preguntas de respuesta estructurada. Los ordenamientos basados en sus resultados son imprecisos y dan lugar a cambios inexplicables de lugar de las escuelas de un año a otro, especialmente las que tienen pocos alumnos. Aunque su propósito de corregir las desigualdades de la situación inicial de los alumnos es correcto, por diversas razones técnicas y prácticas los recientes acercamientos que utilizan modelos de valor agregado tampoco son satisfactorios, al menos por ahora.

Aunque no lo pretendan, los rankings vuelven de alto impacto las pruebas en que se basan, lo que lleva a prácticas que afectan la calidad de la enseñanza, como reducirla al contenido de las pruebas o formas más o menos sutiles de hacer trampa. Los rankings llaman la atención, pero buena parte de los actores, incluyendo medios de comunicación, políticos y empresarios, desconocen sus limitaciones.

Las estrategias de mejora basadas en estímulos económicos y competencia entre escuelas, con base en resultados de pruebas, no tienen en cuenta la peculiaridad de la oferta y la demanda educativas, que no siguen la lógica del mercado.

Muchas personas no tienen conciencia de lo difícil que es tener buenos resultados con alumnos pobres. Es frecuente que dirigentes empresariales vean con simpatía estrategias simplistas de mejora, pensando que las fallas de la escuela pública se podrían corregir fácilmente con escuelas privadas como las que atienden a sus hijos, ignorando que solo una minoría, generalmente privilegiada, asiste a ellas. Por eso abundan las opiniones de que bastará con aplicar pruebas masivamente y tomar medidas correctivas simples para que la calidad mejore sustancialmente. En otras palabras, las estrategias simplistas de mejora parten de un supuesto falso: que hacer buena educación es relativamente sencillo en cualquier contexto.

La proliferación de rankings y su excesivo peso en las políticas públicas sobre educación están trayendo ya consecuencias negativas serias en muchos países:

- la banalización del debate sobre la calidad educativa, reducido a superficiales debates de los ordenamientos, olvidando la complejidad del tema;

- la mercadotecnia engañosa de las escuelas, sobre todo en el sector privado, que buscan atraer alumnos con base en esos ordenamientos;

- el empobrecimiento del currículo, por la tendencia a enseñar para la pruebas, descuidando aspectos que no serán evaluados, aunque sean importantes;

- el cansancio y desaliento en aquellas escuelas que, pese a sus esfuerzos, no consiguen resultados comparables con los de aquellas cuyos alumnos tienen condiciones más favorables, y la actitud negativa de los alumnos frente a una educación centrada en prepararlos para las pruebas; y

- el empobrecimiento de las políticas públicas, que tienden a buscar soluciones fáciles a problemas complejos, descuidando otros aspectos fundamentales, en particular la equidad.

Lo anterior no debe ser entendido como un rechazo a todo tipo de difusión de resultados. Por el contrario: se considera expresamente que la difusión es un elemento fundamental para que la evaluación contribuya a mejorar la calidad educativa. Pero una buena difusión debe tener en cuenta los alcances y los límites de cada evaluación. En el caso de las evaluaciones en gran escala, como se ha señalado, sus límites las hacen inapropiadas para sustentar decisiones que afecten en lo individual a personas o instituciones, ya que estas decisiones deben basarse en información que tenga características de exhaustividad y precisión que no pueden alcanzarse en gran escala y requieren de acercamientos intensivos más completos.

En particular, la postura que se sostiene en este documento no coincide con la de quienes consideran injusto en sí mismo utilizar las mismas pruebas para evaluar a estudiantes de diferentes contextos. Algunas personas sostienen que es injusto usar una prueba para medir el aprendizaje de alumnos cuyos padres tienen baja escolaridad y asisten a una escuela que tiene recursos limitados, y aplicar la misma vara a chicos de familias acomodadas que van a escuelas privadas. Sin embargo, si

se tiene presente la diferencia que hay entre medición y evaluación se podrá apreciar que esa idea no es correcta.

Medir permite llegar a afirmaciones en términos de más qué o menos qué, o de cuántas unidades caben en, pero no en términos de bueno o malo. Para llegar a juicios de este último tipo, o sea para evaluar, es necesario contrastar el resultado de la medición con un referente no empírico, sino normativo, que establece lo que es deseable, en comparación con lo cual podemos llegar a la conclusión de que una realidad es satisfactoria o no, aceptable o inaceptable.

Esta distinción permite entender que una evaluación del nivel de aprendizaje que alcanzan los alumnos de un sistema educativo debe necesariamente utilizar los mismos instrumentos —la misma vara— para medir a todos los alumnos, pero que para llegar al juicio de valor sobre la situación de determinados grupos de alumnos podrá emplear referentes distintos y que, además, deberán siempre explorarse las razones de las diferencias de los rendimientos de unos alumnos y otros, para evitar sacar conclusiones que, esas sí, pueden ser injustas.

Un ejemplo para ilustrar lo que se viene explicando. La diferente presencia de desnutrición y otras circunstancias en los hogares de niños de distinto nivel socioeconómico, además de incidir en su rendimiento escolar, se refleja también en diferencias en su estatura promedio. No por ello deberán emplearse escalas de medición diferentes para registrar la estatura de unos niños y otros en un estudio sobre esa variable. Para tener mediciones comparables es fundamental, desde luego, utilizar instrumentos iguales.

Para llegar a juicios de valor sobre la situación de ciertos niños puede ser razonable, en cambio, utilizar referentes distintos. El juicio de valor depende del propósito de la evaluación; tener cierta estatura —digamos 160 centímetros— puede dar lugar al juicio de que se trata de un sujeto normal, de uno excepcionalmente alto o uno de estatura insuficiente, si en el primer caso se trata de una valoración general del estado de salud de chicos de 3° de secundaria, en el segundo de uno de niñas de 6° de primaria y en el tercero de seleccionar a los integrantes del equipo de básquetbol de un bachillerato.

De manera análoga, un valor de 500 puntos —igual a la media nacional de una prueba de competencia lectora o matemática que se haya estandarizado en esa forma— podrá ser interpretado como normal, sobresaliente o demasiado bajo, dependiendo de que se trate de un alumno de condiciones familiares y escolares cercanas también al promedio nacional, de uno de condiciones particularmente desfavorables o de uno de un medio claramente privilegiado. Lo que hace cambiar el juicio es el punto de referencia, pero la vara de medir tiene que ser la misma, pues de lo contrario toda comparación pierde sentido.

Lo injusto no es medir la estatura de todos los niños, o su rendimiento en matemáticas o lectura con la misma vara —suponiendo que esté bien construida, lo que en el segundo caso no es sencillo—, sino utilizar referentes inapropiados y, sobre todo, llegar a la conclusión de que la menor estatura o el menor desempeño promedio de los

niños de contexto vulnerable es algo reprochable por ser imputable a su menor esfuerzo, o que puede ser imputado a fallas de su maestro.