Revista Iberoamericana de Evaluación Educativa

EVALUACIÓN DE LOS PROCESOS DE EVALUACIÓN DEL SISTEMA EDUCATIVO 1950-2008

Ernesto Schiefelbein y Paulina Schiefelbein

Revista Iberoamericana de Evaluación Educativa 2008 - Volumen 1, Número 1

http://www.rinace.net/riee/numeros/vol1-num1/art3.pdf

El crecimiento de los sistemas educativos, en las últimas seis décadas, ha estado asociado a profundas revisiones (y significativas mejoras) en la manera de evaluarlos. Tiene especial importancia revisar los cambios producidos en tres dimensiones diferentes (pero complementarias), ya que si bien su impacto final parece ser modesto (McKinsey, 2007:5), permite sugerir nuevas formas de mirar la dinámica instruccional (Ball y Forzani, 2007:530) y, eventualmente, la educacional.

El principal cambio consistió, quizá, en pasar de "contar alumnos en cada nivel" a estimar "niveles de aprendizaje" de contenidos y habilidades o aptitudes. Mientras que hace treinta años cualquier evaluación era considerada como un tabú, en la actualidad la mayoría de países someten a los estudiantes a pruebas desde los primeros años de la escolaridad hasta el final de los estudios universitarios (Schiefelbein y Schiefelbein, 2004). Es importante, además, el haber pasado del "clasificar" a los alumnos (entre los que podían avanzar al siguiente nivel y los que debían repetir), a lograr que cada alumno aprenda al menos el mínimo necesario para avanzar (si tiene la capacidad necesaria, aunque sus conocimientos previos sean limitados). Finalmente, se debe destacar el creciente interés por evaluar no solamente el nivel de "instrucción" (en contenidos y habilidades), sino que también la "formación" (propiamente tal) de la capacidad de tomar decisiones responsables conforme a criterios aceptados.

En este artículo se examina los cambios en cada una de estas tres dimensiones y se comenta la manera en que cada uno de estos cambios ha repercutido en la operación de los sistemas y en los objetivos que privilegian los maestros en el aula. Se examina, además, las brechas que estos cambios han generado entre la formación inicial de los maestros y las demandas sociales de calidad de la educación.

1. Pasar desde las brechas de cobertura a las brechas de calidad

De los indicadores de analfabetismo y cobertura usados en la primera mitad del siglo 20 se ha pasado, a través de etapas graduales, a usar los puntajes en pruebas de rendimiento académico. Las "tasas de analfabetismo" eran calculadas con las declaraciones que los "jefes de hogar" hacían a los encuestadores que visitaban cada casa cuando se realizaba un Censo de Población (esto solía ocurrir cada diez o más años) y, por lo tanto, era una medida que dependía de lo que cada jefe de hogar considerada que era ser analfabeto. Se usaban en los estudios internacionales que ha realizado Unesco (desde sus inicios) para comparar el avance de los países en materia de educación y estimular la ampliación de la cobertura.

Chile redujo su tasa de analfabetismo de 25% en 1930, a 16% en 1960, a un 6% en 1992 y a un 4% en la actualidad (MINEDUC, 2005). Si bien Unesco consideraba alfabeto a alguien que había cursado al menos un cuarto grado de primaria, las declaraciones al Censo de población estaban asociadas a que la persona hubiera estado matriculada en algún momento en una escuela.

En los años 60 la mitad de los Ministerios de Educación de América Latina tenía estimaciones de la matrícula en cada grado y, en muchos casos, disponían de datos por edades calculados en los grandes computadoras (denominadas "mainframe" por estar en una gran caja de metal que ocupaba una enorme habitación). Esto permitía calcular las tasas de escolarización (bruta y neta) y, eventualmente, tasas de repetición y de graduación (Schiefelbein, 1974).

Las tasas de repetición permitían suponer, en la década de los 60, que un 20% de los niños de América Latina (que ingresaban a la escuela) tenía problemas de aprendizaje, pero el problema era considerablemente mayor. En efecto, los datos para calcular esas tasas eran proporcionados por los directores de escuela quienes informaban de los alumnos que repetían y de los que abandonaban la escuela, sin saber que muchos de ellos se matriculaban al año siguiente en otra escuela como repitientes. Esto generaba una subestimación de la tasa de repetición y una sobreestimación de la tasa de deserción. Modelos de simulación permitieron calcular, con mayor precisión, que las tasas de repetición del primer

grado estaban cercanas al 50% (Schiefelbein, 1975). Estos indicadores sugerían la necesidad de estimular el aprendizaje de los niños antes de ingresar a la educación primaria (Schiefelbein *et al.*, 2000).

El desarrollo de los lectores ópticos y programas de computación permitió procesar, a fines de los 50, pruebas objetivas administradas a miles de alumnos. Los investigadores y administradores de la educación de doce países desarrollados acordaron en 1958 hacer comparaciones internacionales masivas (Suter, 2001). Estas comparaciones permitieron estimar las brechas en los niveles de instrucción alcanzados en los países. El momento era oportuno porque muchos países de menor desarrollo estaban completando su capacidad de atender a los alumnos de familias con menos recursos de las zonas urbanas. Estos niños tenían problemas para ser promovidos y se acumulaban los alumnos con extra-edad en los primeros grados del sistema.

2. Del acreditar conocimientos (clasificar alumnos) al optimizar aprendizajes

Hasta los años 60 era parte de la responsabilidad profesional del maestro el promover un alumno al curso siguiente sólo cuando estaba seguro que podía aprender al nivel esperado en ese curso (y no se sentía responsable por los que no alcanzaban ese nivel mínimo). El maestro evaluaba al alumno a través de Pruebas de Conocimientos o Capacidad (preguntas escritas a las que el alumno daba una respuesta escrita) y observaciones personales. El criterio subjetivo de cada maestro determinaba quienes quedaban clasificados para aprobar el curso y quienes debían repetirlo. En Chile sólo era posible comparar los niveles de aprendizaje al rendir la Prueba de Bachillerato (administrada hasta 1966 por la Universidad de Chile) una vez que se completaba el nivel secundario. Pero sólo un 5 a 10% de cada grupo de edad rendía esa prueba.

Entre 1965 y 2001 la IEA realizó mediciones de lo que aprendían los alumnos en matemática (en 1965, 1982, 1995 y 1999); ciencias naturales (1970, 1986, 1995 y 1999); y lectura (1970, 1991 y 2001) lo que permitió estimar las brechas que existían entre países desarrollados y en desarrollo (Suter, 2001). La identificación de las brechas estimuló el análisis de los factores asociados a los rendimientos y, a continuación, se realizaron estudios para identificar estrategias de enseñanza efectivas (Coleman et al, 1966; Schiefelbein y Simmons, 1979; Velez *et al.*, 1992; Schiefelbein, 1994). Este tipo de estudios produce, gradualmente, una separación entre los que ejecutan programas educativos y los que los evalúan (McKinsey, 2007:36).

Los estudios sobre escuelas y maestros efectivos de los años 70 llevaron a medir el logro de conocimientos y habilidades específicas (por ejemplo, capacidad de identificar la idea principal de un texto de 100 palabras) mediante pruebas referidas a criterios específicos (Bond, 1996). La selección de ítems de comprensión de lectura ha permitido constatar que en América Latina la mitad de los niños de tercer o cuarto grado no logra entender los mensajes de textos breves y que esta dificultad estaría asociada a insuficiente fluidez de lectura y a un vocabulario relativamente reducido. En la medida que se han detectado estos problemas, las pruebas nacionales se tienden a concentrar en esclarecer las características del problema y dejan de medir otros aspectos de la lecto-escritura como ortografía, gramática o morfosintaxis.

La evaluación referida a norma (clasificar a los alumnos de acuerdo al puntaje en una prueba) vuelve a tomar importancia cuando se constata que en los establecimientos que atienden a niños de nivel socioeconómico bajo sólo se estudia parte del currículo. Esto hace que esos alumnos tengan bajos puntajes en las pruebas de ingreso a la universidad. Sin embargo, los mejores alumnos de cada curso (el 10% con más alto rendimiento) tienden a lograr excelentes rendimientos (en el siguiente nivel) a pesar de sus bajos puntajes en las pruebas de conocimiento (Gil, 2005). Se recupera —mediante esta manera de

seleccionar a los alumnos— a jóvenes de talento excepcional, que por falta de una instrucción adecuada (en el nivel anterior) quedarían marginados de continuar estudiando en el nivel superior.

Pero los resultados de las pruebas no permiten avances significativos en los aprendizajes, ya que los usuarios enfrentan numerosos obstáculos. Algunos ejemplos permiten ilustrar los tipos de obstáculos: (i) distinguir entre acertar respuestas y saber las respuestas; (ii) diferenciar las probables "causas" de un cambio en una variable independiente, de las asociaciones o correlaciones espurias o (iii) estar alertas frente a procesos de "desorientación" en los medios de prensa, por ejemplo, distinguir entre los aportes positivos y negativos de las Tecnologías de la Información y la Comunicación (TIC). Muchas veces los resultados solo permiten suponer soluciones alternativas (Schiefelbein, 2005a, 2004a).

Los que toman decisiones (que no son expertos en el tema) no reciben el mensaje que menos de la mitad de los alumnos de cuarto grado entienden lo que leen y muchos quedan con la sensación que "unos dos tercios" comprende lo que lee. Esto se genera porque los resultados (para el nivel primario) se suelen difundir en términos de porcentajes "brutos" de respuestas correctas. Si en un grupo de 100 alumnos hay 52 que responden correctamente (tasa "neta" de 52% que refleja los que "saben" la respuesta), los 48 restantes (en promedio, cuando hay 4 alternativas) acertarán doce veces, por lo que la tasa bruta llegará al 64% de aciertos. En el SIMCE 2006 el 61,4% acertó con la idea principal del texto de dificultad "intermedia", pero sólo el 48,7% "sabía" cual era la idea principal (SIMCE, 2007b). Este resultado dista mucho de "un buen rendimiento" donde cada uno de los niños tiene éxito (McKinsey, 2007,37).

También la difusión de resultados suele "desinformar o alejar a la opinión publica de los verdaderos problemas" (Brunner, 2007). Esto ocurre, por ejemplo, cuando los titulares de los medios de prensa dicen que "la educación privada logra 60 puntos más que la educación municipal lo que refleja su mejor calidad". Lo que no mencionan los medios de prensa es que cuando los resultados se controlan por el nivel socioeconómico de la familia no hay diferencias en los puntajes (el nivel alto logra unos 300 puntos, el medio obtiene 250 y el bajo entre 220 y 230). La conclusión es inversa (a lo afirmado en la prensa): no hay diferencia de calidad entre los establecimientos (privados y públicos) y, en cambio, los profesores no están preparados para compensar las diferencias en los niveles iniciales de educación (0 a 6 años).

Los mensajes sobre las TIC suelen desorientar al lector desprevenido. Se señala, por ejemplo, que los nuevos audífonos o los mensajes de texto y fotografía de los teléfonos celulares han elevado la cantidad de "copia" o fraude en las pruebas. El mensaje es que el problema estaría generado por la tecnología. Sin embargo, el verdadero problema está generado por el énfasis en que los alumnos memoricen información en vez de que aprendan a utilizarla. En este caso la tecnología ayuda a encontrar la verdadera solución: que las preguntas impliquen un mejor nivel de aprendizaje (por ejemplo, aplicación, comparación, inferencia o juicio crítico, según la taxonomía de Bloom) que no es posible copiar de un libro o un mensaje de texto. En realidad la mayor parte de las pruebas se podrían hacer con derecho a consultar los libros y apuntes (pruebas con libro abierto).

Finalmente, el uso de la información que proporcionan las evaluaciones, para mejorar la educación, implica un uso profesional del conocimiento acumulado hasta ahora, tomando en cuenta sus diversos niveles de confiabilidad (Ravela, 2005; Chapman y Mahlck, 1993). Por ejemplo, si el problema fundamental es el bajo nivel de comprensión de lectura, será necesario examinar las revisiones de las investigaciones pertinentes (Show *et al.*, 1998; Abadzi, 2006; Oliveira, 2006; Bloom, 1976). En muchos casos usar la información que proporcionan las evaluaciones implicará reanalizar los datos con teoría o modelos más adecuados (por ejemplo, puntajes de un grupo de ítems de comprensión de lectura o nivel de vocabulario) o realizar estudios complementarios de los alumnos (por ejemplo medición de aptitudes y personalidad) o de otros aspectos, por ejemplo, examinar características de los profesores o de sus formadores como maestros (Schiefelbein *et al.*, 2007; Schiefelbein, 2004b; Schiefelbein, 1994). En muchos casos se trata de entender el contexto en que ocurren los procesos y las interacciones entre los

diferentes elementos mediante una o varias visitas de un grupo de especialistas a una escuela específica cada cierto número de meses o años, que puede ser menor para las escuelas con mejores indicadores de rendimiento Esto ha permitido reducir a la mitad el tiempo que se dedica a evaluar en Inglaterra (McKinsey, 2007:37).

Este considerable aumento de los instrumentos disponibles para evaluar y de información acumulada sobre procesos de enseñanza y aprendizaje no debe hacernos olvidar que: (i) evaluamos en función de ciertos objetivos que se consideran importantes y (ii) no siempre es posible medir con precisión los objetivos importantes. Es por esto que se evalúa, fundamentalmente, la instrucción, y pone menos atención a la "educación propiamente tal" (porque no podemos medir con facilidad actitudes y procesos de decisión).

Por ahora todavía es útil la evaluación de niveles de instrucción (tal como antes fue la declaración en los censos o la cobertura), pero no hay duda que se necesita más precisión para medir procesos de formación de personas (Schiefelbein, 2003a). Es por esto que se debe revisar qué es lo que realmente queremos obtener cuando evaluamos. ¿Nos interesa saber más, mejorar el mundo (ciencia o técnica) o desarrollarnos como personas?

3. Del evaluar conocimientos (instruir) a formar personas (educar)

A mediados de los 50 se pensaba que el problema de la educación consistía en construir más escuelas y contratar más profesores (y por eso se evaluaba el sistema con los indicadores comentados en el punto 1). Se esperaba que, eventualmente, los niños que estudiaran en esas escuelas se graduarían con niveles similares a los de los alumnos que se graduaban en los 50. Los Bancos internacionales de fomento prestaron fondos y los países ampliaron su cobertura. Pero la ampliación implicó el ingreso de alumnos cuyas familias quedaban por debajo del nivel socioeconómico promedio del país (hasta los años 50 se atendía sólo a la mitad de los niños en edad escolar de primaria cuyas familias quedaban por encima del nivel socioeconómico promedio del país). Los recién ingresados tenían un menor vocabulario, asistían, en promedio, menos días a clases que los de familias con más recursos —la necesidad de cuidar a hermanos o la casa, el caminar mayores distancias (y a veces con ropa inadecuada), la desnutrición y menor protección de la salud o la desorganización familiar están asociadas a mayores inasistencias a clases— y, en general, sus aprendizajes eran menores que los del grupo anterior.

En las última dos décadas, entonces, el problema ha sido como lograr ciertos niveles mínimos de instrucción, principalmente la capacidad de leer y entender los mensajes fundamentales de un texto y realizar ciertos cálculos aritméticos básicos. La etapa siguiente podría ser la capacidad de usar la información (más que la capacidad de recordar datos o aplicar algoritmos que se mide en la actualidad) y en ese caso se usarían pruebas para medir aptitudes (por ejemplo, pruebas con libro abierto, quizá con el estímulo de la tecnología que facilitad el "copiar" información a distancia mediante audífonos o celulares). Pero esto está muy lejos de un esfuerzo de cambiar lo que pasa en las mentes y corazones de los niños (McKinsey, 2007:5).

En la medida que la educación (y la evaluación) se ha reducido, en gran medida, a una mera instrucción, es necesario recuperar *"un justo equilibrio entre la libertad y la disciplina. . . [en el que] . . . la educación lograda es una formación para el uso correcto de la libertad. Hay que aceptar el peligro de la libertad, pero con la corrección de las ideas y decisiones equivocadas, sin apoyar en los errores o fingir que no los hemos visto o, aún peor, que los compartimos, como si fueran las nuevas fronteras del progreso humano"* (Benedicto XVI, 2008). Esto implica empezar a medir lo que se logra en "formación" (porque de otra manera sólo se pone atención a la "instrucción" por ser lo que se mide).

¿Qué indicadores usar para saber cuál es el nivel actual o el cambio en el uso de la libertad? ¿Queremos medir con precisión o saber la magnitud del problema? ¿Cuál es el costo de usar los indicadores de formación? ¿Es posible grabar clases, discusiones o dramatizaciones --en las cuales los participantes representan teatralmente, el papel del ocupante de un determinado rol en la clase-- y luego re-analizar la actuación de los participantes (un psicólogo puede colaborar en analizar el desempeño de cada participante)? ¿Sería posible usar pruebas psicométricas para evaluar aptitudes o rasgos innatos de los participantes a fin de tratarlos y transformarlos en habilidades o capacidades personales para tomar decisiones razonadas, conforme a la conciencia moral? ¿Hay elementos de selectividad en los instrumentos (que generen nuevos problemas de inequidad)? En todo caso es posible visitar las clases y obtener indicadores de los procesos de enseñanza relacionados con formación (Schiefelbein, 2005b).

Por ahora, cuando la mitad de los niños de cuarto grado de Chile no logra comprender textos simples, las alternativas de evaluación son más simples y se analizan en la última sección de este artículo.

4. Cómo usar hoy en Chile los resultados de las evaluaciones: ¿castigar o ayudar?

Desde fines de los 60 se ha medido el nivel de aprendizaje de los alumnos para elevarlo, pero no ha sido evidente la forma de hacerlo y, de hecho, no ha aumentado. La prueba de 8° grado (1967-1971) se limitó a mostrar a los profesores los tipos de habilidades que debían estimular en sus alumnos. La PER, creada a principios de los '80, buscó determinar el nivel del servicio educativo que se ofrecía y "monitorearlo" con ayuda de los padres, que tratarían de seleccionar las mejores escuelas para sus hijos. El SIMCE midió, a partir de 1988, los rendimientos de las escuelas y trató de identificar factores que pudieran explicar las diferencias y evaluar el impacto de los programas de los municipios y del MINEDUC.

Sin embargo, no hay avances y conviene repensar este poderoso instrumento, que es el SIMCE, si se quiere mejorar los aprendizajes. Un rol alternativo del SIMCE sería, según algunos expertos, ayudar a los profesores a enseñar mejor y revisar cuáles son los temas importantes del currículo, más que el "clasificar" rendimientos. Este segundo rol plantea diferencias nítidas con respecto al primero: ¿Seleccionar o desarrollar? ¿Poner nota o medir logro de meta? ¿Conocer el ranking o el aprendizaje de objetivos específicos (criterio)? ¿Castigar o ayudar? "Son dos funciones diferentes (los especialistas hablan de Evaluación Sumativa y Formativa, respectivamente). Si la primera no elevó hasta ahora los rendimientos, convendría probar la segunda (Schiefelbein, 2003b).

Esto implica fuertes cambios en el SIMCE. En efecto, para "seleccionar" o "castigar" basta calcular un puntaje total que "discrimine entre buenos y malos" y difundirlo públicamente. En este rol es fundamental no divulgar los ítems utilizados para que mantengan su poder de discriminación (no ser conocidos por los profesores y evitar que preparen a sus alumnos para contestarlos mecánicamente). Basta calcular el puntaje de cada escuela para que los padres seleccionen entre escuelas o preparar datos e indicadores globales para que los funcionarios del MINEDUC modifiquen las estrategias.

En la segunda función, en cambio, el usuario de la información del SIMCE es el profesor en su sala de clases. En efecto, para "desarrollar" o "ayudar a mejorar" se necesita entregar información detallada a cada profesor sobre los aspectos en que cada uno de sus alumnos logró los niveles adecuados y aquellas habilidades o conocimientos que todavía no posee o no domina suficientemente. El profesor debe saber lo que contestó cada alumno en la prueba y revisar con cada uno los errores que cometió, hasta que el alumno internalice las deficiencias y reforme adecuadamente sus procesos de pensamiento. Sin información detallada el profesor no puede identificar los aspectos de su enseñanza que debe cambiar o cómo ayudar a cada estudiante... *"En resumen, una redefinición clara del objetivo de la prueba nacional de medición de la calidad de la educación debe orientar la reflexión del grupo de expertos sobre los cambios que debe tener el SIMCE..."*