Criterios de evaluación: bases y perspectivas

Andrés Sánchez Moguel

La evaluación del aprendizaje es una práctica de suma importancia en el ámbito educativo a la que, sin embargo, se le ha dado poco espacio en la reflexión de los estudiosos de este tema en nuestro país. El estado de conocimiento sobre la evaluación del aprendizaje en el II Congreso Nacional de Investigación Educativa (Martínez y cols., 1993) señala que son muy pocos los trabajos realizados sobre este aspecto de la evaluación educativa a lo largo de la década comprendida de 1982 a 1992 (se encontraron únicamente 81 estudios, tesis y artículos, en revisión exhaustiva). Ya García Cortés en 1979 señalaba el problema de la "realización paupérrima de estudios de evaluación educativa en nuestro medio". En el III Congreso Nacional de Investigación Educativa (1995) se sigue viendo la dispersión de la temática, ya que los trabajos presentados son en general técnicos y centrados en problemas relativamente aislados (Sánchez Moguel, 1997).

La falta de interés teórico por el tema y el consiguiente poco espacio dedicado a su análisis en las instituciones formadoras de maestros ha llevado a la generalización de prácticas evaluativas del logro escolar con las siguientes características:

a) Falta de reflexión sobre las razones por las cuales se evalúa, dando prioridad al cumplimiento administrativo sobre la utilidad real de la información.

b) Procedimientos e instrumentos de evaluación poco planeados y mal estructurados.

c) Escaso análisis de lo obtenido en las evaluaciones, priorizando los intereses crediticios ("aprobé o no", "quince alumnos reprobaron", etcétera) por sobre los logros académicos ("he aprendido al cien por ciento este contenido", "ya hay un conocimiento generalizado de esta materia o no", etcétera).

d) Una serie de factores que distorsionan la medición de lo que los estudiantes realmente saben, tales como ligar la conducta a la calificación, las altas posibilidades de fraude o las pruebas que privilegian lo memorístico sobre lo reflexivo.

Por todo lo anterior, es importante regresar siempre a las bases, a las razones por las cuales se hace la evaluación educativa. El objetivo del presente trabajo es analizar algunas cuestiones de fondo acerca de la teoría de la evaluación, en su relación específica con la evaluación del aprendizaje, con la intención de mostrar a los partidarios de las posturas extremistas que las prácticas evaluativas en el aula, en la escuela y en el sistema educativo no tienen por qué ser ciegas recetas que inventó un técnico, y que debe haber una reflexión seria sobre la manera de abordar la tarea de evaluación y de extraer el significado de cada dato obtenido.

La evaluación educativa es una estrategia de recolección de información sobre los diferentes momentos, actores y auxiliares del proceso enseñanza-aprendizaje. Si bien es indispensable que cada profesor haga evaluaciones particulares y a profundidad de estos elementos al interior de su espacio de trabajo, es también necesario contar con perspectivas más generales de la labor académica que sirvan de monitor a la totalidad de la comunidad de las escuelas. Por tanto, la evaluación masiva de estudiantes, la evaluación reflexiva entre los maestros y la evaluación institucional se presentan como prácticas útiles en el acopio de datos globales sobre la situación escolar. Múltiples circunstancias han hecho que en algunas ocasiones la evaluación general se lleve a cabo con una perspectiva de conteo y control que

recaba datos con instrumentos que no han sido depurados, asigna calificativos por simple tradición numérica,1 y genera listados llenos de cifras a los que no se da un uso en el perfeccionamiento del esquema educativo.

La intención de quienes realizan el proceso de evaluación y quienes lo promueven es decisiva en el énfasis que se va a dar a unos u otros elementos del sistema de evaluación que se genere. Así, ante el único interés de cumplir de la manera más eficiente posible con una exigencia administrativa o estatutaria, los evaluadores prefieren hacer instrumentos que estén listos para ser aplicados en muy poco tiempo, que sean lo más económicos posible en tiempo, dinero y esfuerzo, así como que el análisis de los datos resultantes de las aplicaciones sea rápido y no evidencie las carencias de sus instrumentos, aunque este análisis no tenga un uso fuera de los expedientes. En cambio, ante el interés de obtener información útil en el análisis de la situación educativa, las preferencias tienden a desarrollar instrumentos con un nivel suficiente de validez, confiabilidad y pertinencia que lleven a obtener datos adecuados para el propósito de conocer los resultados escolares, permitiéndose así hacer mayores inversiones de recursos que en otros modelos.

Varios puntos de decisión se plantean en este proceso: ¿quién debe decidir las áreas a evaluar, los contenidos de las áreas a evaluar y los métodos de evaluación?, ¿quién debe crear los instrumentos de evaluación?, ¿qué características deben tener los ítemes de los instrumentos y/o métodos de evaluación?, ¿en qué momento se pueden considerar adecuados los instrumentos y/o métodos de evaluación?, ¿qué criterios deben tomarse en cuenta para analizar la información obtenida?, ¿qué segmentos y agrupaciones de la información obtenida son más útiles, y para quién? La pertinencia y la utilidad del proceso dependen de que se den respuestas razonables a estas preguntas, en los momentos oportunos, perfilando un sistema.

Existe una serie de cuestiones con respecto a la forma de abordar el trabajo de la evaluación masiva que ha sido resuelta con base en criterios poco claros, e incluso sin siquiera considerar los problemas teóricos y de aplicación, planteando las decisiones en términos meramente técnicos y/o administrativos. García Cortés (1979) explica la gran importancia que tiene determinar, para cada caso específico, las respuestas a para qué evaluar y qué evaluar. Responder estas dos preguntas señala criterios que generalmente sirven de gran ayuda para tomar decisiones sobre la manera de operar un programa de evaluación.

Es conocido el hecho de que un instrumento y/o método de evaluación (desde una regla hasta un electroencefalógrafo) que va a ser utilizado en repetidas ocasiones, para sacar conclusiones al hacer comparaciones debe cumplir ciertos criterios de confiabilidad y validez, así como ser pertinente.

La confiabilidad se refiere a la estabilidad del instrumento a través del tiempo y de las muestras. Sabemos que las condiciones y cualidades de los actores educativos son dinámicas, así que esta primera definición no parece ser muy útil en el ámbito de la escuela. Una segunda aproximación refiere que la medida confiable es aquella que se encuentra libre de error. Sin embargo, aunque esto parece ser suficientemente exacto (nunca totalmente), en las ciencias naturales (por ejemplo, en la medición del contenido de sodio en un compuesto) y en las ciencias sociales es muy ingenuo pensar en alcanzar la exactitud (puede incluso plantearse la duda de la posibilidad o la necesidad de ella en el plano filosófico). Una definición que nos parece más viable para la tarea que nos ocupa es la de considerar semejante a lo que es semejante, y diferente a lo que lo es, lo cual acerca la noción cuantitativa de confiabilidad a la noción cualitativa de imparcialidad (Fernández Ríos, 1994).

Los estándares para la evaluación educativa y psicológica por medio de pruebas (apa, 1985) señalan que "la validez es la consideración más importante en la evaluación por medio de pruebas. El concepto se refiere a la pertinencia, significación y utilidad de las inferencias específicas que se hagan de los puntajes de una prueba". Es muy difundida la definición básica de validez en instrumentos de evaluación que indica que éstos son válidos cuando miden lo que pretenden medir (Magnusson, 1975). Sin embargo, este concepto de validez aparentemente tan simple se encuentra en el centro de una polémica que aún actualmente se lleva a cabo. Gray (1997), haciendo una pequeña revisión, señala que:

…en 1949 Cronbach declaró que la definición de validez como "la extensión con que una prueba mide lo que pretende medir" era comúnmente aceptada, aunque él prefería una ligera modificación: una prueba es válida en el grado en que sabemos qué mide o predice. Cureton (1951) provee una definición similar: la cuestión esencial de la validez en las pruebas es qué tan bien realizan la tarea para la cual se les está usando. La validez es definida entonces en términos de la correlación entre los puntajes de una prueba y los "verdaderos" puntajes del criterio. La perdurable definición de Anastasi (usada desde 1954 hasta 1997), "la validez es qué mide una prueba y qué tan bien lo hace", es también citada ampliamente.

Gray (op. cit.) señala también que, aunque Cronbach tendió a evitar redefinir el término surgido en 1949, en 1971 hizo un comentario que reavivó la controversia: "validación es el proceso de examinar la precisión de una predicción o inferencia específica hecha a partir de los puntajes de una prueba", o bien, como señalan otros autores, "la validez se refiere no a las puntuaciones o datos en sí mismos, sino a las inferencias que se hagan a partir de ellos bajo determinadas circunstancias" (Cronbach, Vernon, cit. en Silva y Martorell, 1991); "lo que se valida no es el instrumento, sino la interpretación de los datos obtenidos por medio de un procedimiento especificado" (Aragón, 1990); "la validez depende de la 'adecuación y pertinencia de inferencias y acciones' basadas en los resultados de la evaluación" (Messick, 1989, en Linn y Baker, 1996).

Finalmente, es importante señalar que, aunque muchos autores (Rudner, 1993; Niemi, 1996; Aragón, op. cit.; Tourón, 1989; Burns, 1996; gao, 1991) reportan al menos tres tipos "clásicos" de validez, actualmente existe una tendencia a considerar un tipo único de validez (Gray, op. cit.; Silva y Martorell, op. cit., quienes incluso sugieren que el concepto de confiabilidad también es mucho más cercano al de validez de lo que se ha pensado), del cual, eso sí, se obtienen distintos tipos de evidencias:

Se ha sugerido que la validez de constructo abarca tanto a la validez de criterio como a la de contenido. Sheperd anotó que la validez de constructo incluye los requisitos teóricos y empíricos de la validez de contenido y de criterio. Anastasi (1986) coincide en que la validez de constructo subsume los requisitos de la validez de contenido y de criterio. (Stapleton, 1997)

En resumen, debemos considerar como una cualidad primordial de las pruebas la posibilidad de extraer de manera correcta y verdadera el significado de sus puntajes. Dado que esto no depende sólo de la prueba sino también de las circunstancias de aplicación y los objetivos de la misma, diferentes aspectos de esta cualidad pueden ser considerados. Aunque esto puede parecer sencillo cuando los instrumentos de medición son muy cercanos a la realidad física, la tarea se complejiza conforme el objeto de evaluación se vuelve abstracto o difícil de observar directamente. Tal es el caso de la evaluación del aprendizaje.

Existen tres puntos de especial importancia en cuanto a la pertinencia de un procedimiento de evaluación:

1. Que el tipo de información arrojada sea realmente un indicador útil sobre los conocimientos y/o habilidades de la población.

2. Que existan criterios fundamentados para interpretar las cifras obtenidas en la examinación masiva.

3. Que la información obtenida llegue a los destinatarios que pueden darle utilidad, es decir, los profesores, planificadores académicos al interior de la escuela y los propios estudiantes.

De lo anterior se desarrollan los siguientes puntos:

• Que el tipo de información arrojada sea realmente un indicador útil sobre los conocimientos y/o habilidades de la población. Existe una discusión importante con respecto a los instrumentos de evaluación que se utilizan en educación. En realidad, el origen de la discusión está en el pseudoproblema de lo cuantitativo versus lo cualitativo. Algunos autores, como Díaz Barriga (1982), señalan que la evaluación no debe hacer uso de la tecnología de medición generada por la psicometría y perfeccionada constantemente pues "se minimiza tanto el proceso mismo de la evaluación del aprendizaje como la noción de aprendizaje y la de docencia."; otros plantean problemas técnicos en el uso de ciertos tipos de evaluación "objetiva", por ejemplo que sólo se mide lo que el alumno memoriza, o la posibilidad de acertar por azar (Fermín, 1971); finalmente, otros autores, reconociendo los problemas de "el hiato indudable entre la medida y lo que pretendemos medir" y "el uso de la medición en la evaluación educativa" (Tourón, 1989) confían, sin embargo, en el uso del método científico para la valoración escolar y generan estrategias cada vez más refinadas para salvar los problemas mencionados (Tourón, 1989; Tirado y Serrano, 1989; Rodríguez y García, 1982). Consideramos importante rescatar nociones de cada uno de estos planteamientos, que equilibren una práctica evaluativa eficaz, eficiente y útil. Así, creemos que el diseño de un instrumento de evaluación debe estar firmemente enraizado en una reflexión del para qué y el qué evaluar, de tal modo que si una técnica y el qué evaluar se revelan incompatibles, debe ser la técnica la que cambie. También que el trabajo teórico cuidadoso con academias de evaluación puede llevar a que éstas diseñen métodos e instrumentos de evaluación adecuados para la intención, lugar y momento específicos en que se reflexione como grupo. Y que la información obtenida con métodos e instrumentos puede ser analizada y divulgada de manera útil para la toma de decisiones en la institución educativa.

• Que existan criterios fundamentados para interpretar las cifras obtenidas en la examinación masiva. Una problemática común entre los que atacan el problema de la evaluación desde un punto de vista social y/o filosófico, que en cambio es poco tocado por quienes tienen el punto de vista únicamente técnico, es el criterio de pase-reprobación en el caso de evaluaciones con fines de acreditación, o el criterio de "aceptabilidad-inaceptabilidad" en el de evaluaciones para la toma de decisiones. Sabemos que existen en este sentido juicios "por criterio" y juicios "por norma". En los primeros, se establece de antemano el mínimo aceptable, que depende de una discusión teórica de lo que se va a evaluar, y en los segundos se juzga cada caso individual con base en la cercanía o lejanía que tenga con la media (por ejemplo, número de desviaciones estándar), y el sentido de esta distancia (positivo o negativo). En el papel, estos criterios pueden parecer fáciles de aplicar, pero en la práctica vale la pena reflexionar profundamente en los motivos y las consecuencias de permitir, por ejemplo, que sean acreditados estudiantes de medicina con

calificaciones apenas pasables, además de relativas (dado que ante el examen de una escuela podrían obtener altas calificaciones y ante el de otra podrían ser bajas). En efecto, no hay una estandarización en la dificultad que deben tener este tipo de pruebas, ni normas o consejos de uso generalizado para establecer los criterios. Por todo ello, el conjunto de la sociedad escolar debe dedicar tiempo a la reflexión de este problema, aterrizándolo en programas concretos en los que se trabaje, y tomando decisiones con respecto a los criterios a emplear en ellos.

La relatividad llegó a la física —una de las ciencias más duras y clásicas— hace unos ochenta años; tal vez ya es tiempo de que llegue a la educación: no existen criterios ni fórmulas universales para llevar a cabo las tareas evaluativas, ni deben existir. Cada sociedad escolar debe definir los propios. "Las interpretaciones válidas del significado y la verdad son hechas por gente que comparte decisiones y las consecuencias de las decisiones", escribe Steinar Kvale a propósito del conocimiento (según traducción inédita de Carrascosa). Estos términos, llevados a la evaluación educativa, implican el compromiso y la reflexión de todos los participantes en el proceso de la educación.

• Que la información obtenida llegue a los destinatarios que pueden darle utilidad. El último problema que planteamos para reflexionar en cuanto a la pertinencia de la evaluación, es el de decidir la manera de presentar la información obtenida y el análisis realizado con base en la aplicación de los métodos e instrumentos, así como los modos e instancias de distribución de estos datos. Consideramos útil discutir de antemano estos elementos, y evaluar la certeza de nuestras decisiones luego de cada experiencia de divulgación, mejorando sucesivamente las estrategias de difusión con base en las observaciones que se hagan. También consideramos útil consignar el proceso de búsqueda de las mejores estrategias en escritos que puedan ser de utilidad a otros en su práctica evaluativa.

Conclusiones

Con base en lo reflexionado, se infiere que la evaluación de lo educativo es una tarea fundamental, por su función de retroalimentación del sistema y sus subsistemas. Pasaron ya los tiempos de decidir entre una evaluación cuantitativa y una cualitativa, aunque es cierto que subsiste el problema técnico de que algunos tipos de evaluación, por su naturaleza, tienden a pertenecer mayormente a uno de estos dos campos.

La evaluación de lo educativo debe ser llevada a cabo por la comunidad. Debe haber participación de los actores educativos en las diferentes fases de la evaluación, principalmente en las de fundamento (cuando se establecen los criterios, con base en valores reconocidos por el grupo) y en las de retroalimentación propiamente dicha. Una cultura de evaluación no significa una época de terror, de premios y castigos basados en procesos desconocidos que asignan números bajo reglas cabalísticas oscurísimas: ésa es la cultura de la zanahoria y el palo para hacer andar al "motor ecológico". En una cultura de evaluación hay un interés de los participantes del proceso educativo por conocer el desempeño personal y grupal para analizar lo alcanzado y dirigir esfuerzos con conocimiento de causa que aumenten las probabilidades de éxito, y hay también un esfuerzo sostenido por revisar y mejorar constantemente los medios por los que se obtiene la información que sirve de base para los análisis.

Nota

1 Nos referimos especialmente a la conocida "escala de cero a diez", en que seis o más significa "aprobado", es decir, adecuado, y cinco o menos significa

"reprobado", es decir, inadecuado. Del mismo modo podemos hablar del sistema na-mb.

Bibliografía

American Psychological Association, Standards for Educational and Psychological Testing, Washington, D.C., apa, 1985.

Aragón, B. L., Elaboración de un instrumento de evaluación conductual, con validez de contenido y de tratamiento, para niños disléxicos, tesis de grado, enep-Iztacala, unam, 1990.

Burns, W. C., "Content Validity, Face Validity and Quantitative Face Validity", en R. S. Barrett (ed.), Fair employment strategies in human resource management, Quorum Books, 1996.

Díaz Barriga, A., "Tesis para una teoría de la evaluación y sus derivaciones en la docencia", Perfiles educativos, núm. 15, enero-marzo, cise-unam, 1982.

Fermín, M., La evaluación, los exámenes, las calificaciones, Kapelusz, 1971.

Fernández Ríos, L. F., Manual de psicología preventiva: teoría y práctica, Siglo xxi, Madrid, 1994.

gao, Designing Evaluations, United States General Accounting Office, 1991.

García Cortés, F., "La evaluación en la educación", Perfiles educativos, núm. 3, enero-marzo, cise-unam, 1979.

Gray, B. T., "Controversies Regarding the Nature of Score Validity: Still Crazy After All These Years", presentado en la reunión anual de la Southwest Educational Research Association, Austin, enero de 1997.

Kvale, Steinar, "Psicología posmoderna: ¿Una contradicción de términos?" (no publicado), traducción de C. Carrascosa.

Linn, R. L. y E. L. Baker, Assessing the validity of the National Assessment of Educational Progress: NAEP technical review panel white paper, U. S. Department of Education, 1996.

Magnusson, D., Teoría de los tests, Trillas, México, 1975.

Martínez, F. F., G. Fuentes Trejo, B. Cepeda Hinojosa y R. Burgos Fajardo, Estado de conocimiento 8: evaluación del aprendizaje, Comité Organizador del Segundo Congreso Nacional de Investigación Educativa/Sindicato Nacional de Trabajadores de la Educación, 1993.

Niemi, D., Instructional influences on content area explanations and representational knowledge: evidence for the construct validity of measures of principled understanding, National Center for Research on Evaluation, Standards, and Student Testing, 1996.

Rodríguez Cruz, H. M. y E. García González, Evaluación en el aula, Trillas, 1982.

Rudner, L. M. Test Evaluation, eric/ae, 1993 (http://136.242.172.58/intass.htm).

Sánchez Moguel, A., "Evaluación de la educación. Introducción", en Ángel Díaz Barriga (coord.), Currículum, evaluación y planeación educativa, comie, cesu, enep-Iztacala, 1997.

Silva, F. y C. Martorell, "Evaluación conductual y evaluación tradicional: la cuestión psicométrica", en: V. E. Caballo (ed.), Manual de técnicas de terapia y modificación de conducta, Siglo xxi, Madrid, 1991.

Stapleton, C. D., "Basic Concepts in Exploratory Factor Analysis (efa) as a Tool to Evaluate Score Validity: A Right-Brained Approach", presentado en la reunión anual de la Southwest Educational Research Association, Austin, enero de 1997.

Tirado Segura, F. y V. Serrano Carrillo, "En torno a la calidad de la educación pública y privada en México", Ciencia y Desarrollo, vol. xv, núm. 85, Conacyt marzo-abril de 1989.

Tourón, J., "La validación de constructo: su aplicación al ceed (Cuestionario para la evaluación de la eficacia docente)", Bordón, vol. 41 (3-4), 1989.