LA EVALUACIÓN EN EL AULA*

Lorrie A. Shepard
Universidad de Colorado, Campus Boulder

Capítulo 17 de la obra Educational Measurement
(4ª Edición) Editado por Robert L. Brennan
ACE/ Praeger Westport. 2006
pp. 623-646.

*Estoy muy agradecida con Rick Stiggins y Mark Wilson por los alentadores y estimulantes análisis de los borradores de este capítulo

LA EVALUACIÓN EN EL AULA

Coordinación Editorial:
Miguel Á. Aguilar R.
Diana L. Flores Vázquez

Diseño y Formación:
Juan Cristóbal Ramírez Peraza
Irma Tapia Covarrubias

Instituto Nacional para la Evaluación del Educación
José Ma. Velasco 101, Col. San José Insurgentes,
Delegación Benito Juárez, C. P. 03900, México D. F.

Traducción:
Martha Domís para el Instituto Nacional para la Evaluación de la Educación

Impreso en México

CONTENIDO

PRESENTACIÓN

Con este volumen, el Instituto Nacional para la Evaluación de la Educación (INEE) inicia una nueva serie de publicaciones, que difundirá textos relevantes sobre la evaluación educativa creados por la pluma de autores externos al Instituto, o que no caben dentro de las otras series que comprende su programa editorial. En el caso de obras de autores externos, además de su interés, se considerará la dificultad de acceder a ellas por los lectores mexicanos, debido a no estar publicadas en español u otras razones. Este fue el caso de la obra *Learning divides*, del investigador canadiense Jon Douglas Willms, que el INEE hizo traducir y publicó con el debido permiso del editor original, el Instituto de Estadísticas de la UNESCO, con el título *Las brechas de aprendizaje*.

La obra que se difunde ahora, de la profesora Lorrie A. Shepard, de la Universidad de Colorado en Boulder, es un texto fundamental sobre un tema que el INEE considera de gran interés para toda persona interesada en la evaluación, especialmente para los que en el momento actual sienten preocupación por el enfoque que se está dando a la evaluación en gran escala en nuestro país.

Convencido como lo está del potencial positivo de la evaluación en gran escala, el INEE comparte la preocupación de no pocas personas del medio educativo, en el sentido de que ciertos usos de ese tipo de evaluaciones pueden tener también serias consecuencias negativas. Por ello, es de gran importancia reflexionar seriamente sobre los alcances y limitaciones de dichas evaluaciones, así como sobre la necesidad de que se fortalezcan paralelamente las evaluaciones a cargo de los maestros, de manera que la combinación de unas y otras contribuya verdaderamente a avanzar en la dirección que a todos nos interesa: la de la mejora real y profunda de la educación que nuestras escuelas ofrecen a los niños y niñas de México.

La obra que se presenta es un texto de primera importancia en relación con estos temas, y por ello el INEE la pone al alcance de los maestros de nuestro sistema educativo, y de todas las personas interesadas por la calidad educativa. En los párrafos siguientes se desarrollan con mayor amplitud las preocupaciones que nos han llevado a difundirla.

En su forma más conocida, la evaluación educativa no es algo reciente. La tarea del maestro, en su interacción cotidiana con los alumnos, ha incluido siempre, como una dimensión fundamental, el evaluar los avances de cada uno. En las formas tradicionales de enseñanza que prevalecieron hasta bien entrado el siglo XIX, cuando surgieron los sistemas educativos de concepción moderna con los que estamos familiarizados –con cobertura que tendía a ser universal y, por ello, con muchos alumnos, organizados en grados de edad y avance similar– la tarea de los maestros era más de evaluación que de docencia. En las escuelitas en que un *dómine* atendía a una docena de chicos de distintas edades y niveles, para enseñarles a leer, escribir, contar y rezar, la lección magistral estaba ausente; el trabajo del maestro se limitaba fundamentalmente a *tomar la lección* a cada alumno, indicándole, en función de su avance, la siguiente tarea.

La que sí es bastante reciente es *la evaluación en gran escala*, la aplicación estandarizada de

pruebas a grandes números de alumnos, para apreciar el nivel de aprendizaje que se alcanza en el sistema educativo de todo un país, región o distrito, ante la imposibilidad de agregar las evaluaciones que hacen los maestros, siempre ligadas al contexto en que trabaja cada uno.

Con antecedentes que se remontan al final del siglo XIX, las evaluaciones en gran escala se desarrollaron en los Estados Unidos, durante la primera mitad del siglo XX, adquirieron importancia a lo largo de su segunda mitad y se extendieron a la mayor parte de los países del mundo en las dos o tres últimas décadas. Además de evaluaciones nacionales, se desarrollaron proyectos internacionales que hoy atraen poderosamente la atención cada vez que se difunden sus resultados. Las más conocidas son las pruebas del Proyecto para la Evaluación Internacional de los Estudiantes (PISA, por sus siglas en inglés), de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), cuyo desarrollo comenzó a planearse en 1995, y se aplican cada tres años desde el 2000.

Más de tres décadas antes, en 1958, comenzaba a gestarse la Asociación Internacional de Evaluación del Rendimiento Académico (IEA, por sus siglas en inglés), con la planeación del Primer Estudio Internacional sobre Matemáticas que se llevó a cabo en la década de 1960, del que se deriva el Estudio de Tendencias en Matemáticas y Ciencias (TIMSS) una de las evaluaciones vigentes más importantes. En México, el desarrollo de pruebas en gran escala para educación básica dio inicio desde la década de 1970, y se desarrolló sobre todo a partir de la de 1990, con las pruebas para evaluar el Factor Aprovechamiento Escolar del Programa de Carrera Magisterial. La tendencia se acentuó en la última década, con las pruebas del Instituto Nacional para la Evaluación de la Educación (INEE), a partir de 2003, y las pruebas censales de la Secretaría de Educación Pública, desde 2006.

Las pruebas en gran escala pueden ser una herramienta valiosa para apoyar los esfuerzos de mejora de la calidad educativa, si se las ve como un complemento de las evaluaciones a cargo de los maestros, insustituibles para evaluar de manera detallada todos los aspectos del currículo y para hacerlo de manera que puedan ofrecer a cada alumno la retroalimentación precisa sobre sus puntos fuertes y débiles, esencial para mejorar el aprendizaje. Es preciso, sin embargo, advertir sobre un riesgo que no se puede ignorar: el peligro de que las pruebas en gran escala produzcan consecuencias negativas para la calidad educativa, si se les comprende y utiliza mal.

En efecto: para poder dar resultados confiables de los niveles de aprendizaje que alcanzan muchos miles de alumnos, una prueba estandarizada necesariamente tiene que reducirse a la medición de un número relativamente mínimo de temas, y debe hacerlo mediante preguntas que no pueden atender los aspectos más complejos de las competencias que pretende desarrollar la escuela. Por ello, hay que reiterar que las pruebas en gran escala no pueden sustituir el trabajo de evaluación de los maestros, el único que puede atender con precisión los aspectos más complejos de la enseñanza y el aprendizaje, y hacerlo de modo que se brinde retroalimentación detallada y oportuna a cada alumno.

Si no se entienden bien los alcances y límites de los resultados de las pruebas en gran escala, es fácil que se usen en forma inapropiada. El peligro más claro es la tendencia a tomar como referente para la tarea docente el contenido de las pruebas y no el de los programas de estudio, enseñando para las pruebas, por la visibilidad de sus resultados. Con ello la tarea de la escuela se empobrece, al descuidar aspectos esenciales que no evalúan las pruebas en gran escala, como la expresión escrita y oral, la formación de actitudes y valores, la educación artística, e incluso los niveles cognitivos más complejos de las áreas tradicionalmente cubiertas de Lectura, Matemáticas y Ciencias.

Otros ejemplos de manejo inapropiado de los resultados de las evaluaciones en gran escala son la asignación de estímulos a los docentes o la elaboración de ordenamientos simples de escuelas,

supuestamente en función de la calidad de unos y otras, sin tener en cuenta los numerosos factores que inciden en los resultados de los alumnos en las pruebas ni tener en cuenta las limitaciones de éstas. Este tipo de errores, además, produce un explicable rechazo de toda evaluación en gran escala por parte de muchos maestros, que perciben sus graves consecuencias para la educación. Por ello, conviene reiterar que la evaluación en gran escala puede ser muy valiosa para la mejora de la calidad, a condición de entenderla y usarla viéndola como complemento del trabajo del maestro, y no como sustituto del mismo. Para eso es necesario que los maestros entiendan bien los alcances y límites de la evaluación en gran escala, y los de la evaluación que ellos llevan cabo.

En México la formación inicial que reciben muchos maestros no los prepara bien ni para una cosa ni para la otra. Un indicio de ello es la solicitud que el INEE suele recibir, de maestros, supervisores y directivos de escuelas normales, para impartir talleres de elaboración de reactivos de opción múltiple, gracias a lo cual se espera que mejore la calidad de las evaluaciones que deben hacer los maestros. En el Instituto hemos mantenido la posición anterior sobre los alcances y límites de la evaluación en gran escala, y sobre la necesidad de verla como complementaria de la evaluación a cargo de los maestros, y nos enfrentamos permanentemente con una dificultad considerable para conseguir que esta postura sea comprendida no sólo por la sociedad en general, sino también por las autoridades educativas y por los maestros, como muestran las demandas a las que alude el párrafo.

Si se comprenden las características de las evaluaciones en gran escala y en aula, se entenderá también que las preguntas de opción múltiple son esenciales en las primeras, pero que las segundas, las evaluaciones que los maestros llevan a cabo, pueden utilizar acercamientos diferentes y mejores para evaluar aspectos finos y complejos, los cuales difícilmente se pueden atender en gran escala, pero que en el ámbito del aula es posible emplear. Las preguntas de opción múltiple pueden ser usadas también en el aula, y son adecuadas para evaluar algunos aspectos del aprendizaje, pero otros deben valorarse de formas distintas, como mediante la producción de textos amplios, la realización de ejercicios en vivo, la observación del trabajo individual y grupal de los alumnos, entre otros.

La expresión *evaluación en aula (classroom assessment)* se refiere a este tipo de acercamientos. Es importante que escuelas normales y programas de actualización de maestros en servicio presten la atención que merecen a estos enfoques, relativamente recientes y poco conocidos en nuestro medio.

Para contribuir al desarrollo de estas innovadoras ideas, el INEE hizo las gestiones necesarias para difundir en español el texto siguiente, de una de sus principales defensoras. Al invitar a leerlo y reflexionar detenidamente sobre su contenido, expreso el deseo de que la evaluación educativa en México se desarrolle combinando el avance técnico de las pruebas en gran escala, con un uso de sus resultados que no ignore sus límites y un avance substancial de la evaluación en aula, a cargo de los maestros. Así, y sólo así, la evaluación podrá contribuir realmente a la mejora educativa.

Felipe Martínez Rizo
Julio de 2008.

La evaluación en el aula

El modelo de evaluación en el aula que se explica detalladamente en este trabajo es muy diferente del modelo de pruebas y mediciones que predominó en aulas y escuelas durante el siglo pasado. En los primeros años del siglo XX, los expertos en mediciones creían que podían usarse pruebas nuevas y objetivas para estudiar y mejorar los resultados de la educación, así como para encargarse del diagnóstico y la colocación de estudiantes de acuerdo con sus necesidades de aprendizaje (Symonds, 1927; Thorndike, 1913). El punto de vista prevaleciente fue que los expertos debían elaborar pruebas estandarizadas que los docentes utilizarían con objeto de incrementar la precisión en su toma de decisiones. Además, los expertos en mediciones empezaron a enseñar a los maestros cómo hacer sus propias pruebas siguiendo principios científicos de medición. En aquellos primeros años, se desarrolló un sistema para los libros de texto de mediciones con el fin de enseñar a los maestros cuestiones sobre la validez y la confiabilidad (utilizando representaciones en su mayor parte cuantitativas), la elaboración de pruebas, los formatos y el análisis de reactivos, así como análisis estadísticos de los resultados de las pruebas. Este sistema —que consistía casi exclusivamente en pruebas formales, cuestionarios y calificaciones— ha seguido siendo el modelo de los libros de texto hasta el día de hoy.

En contraste con este modelo técnico y cuantitativo, existe un punto de vista diferente de la evaluación en el aula, que se desarrolló a fines del siglo XX y que busca lograr que, en mucho mayor medida, el estudiante alcance un entendimiento; asimismo, busca obtener el uso formativo de la evaluación como parte del proceso de aprendizaje (Black, y Wiliam, 1998; Gipps, 1999; Shepard, 2000). A principios de la década de los ochenta, el interés en reformar la práctica de la evaluación se vio acuciado por un uso mayor de pruebas estandarizadas, cuyo propósito era la responsabilización*, y por una evidencia cada vez mayor de que los formatos estrechos de pruebas tenían un efecto perjudicial en la calidad de la enseñanza y el aprendizaje de los estudiantes (Resnick y Resnick, 1992; U.S. Congress, *Office of Technology Assessment*, 1992). Adelantándose a los expertos en mediciones, los especialistas en las materias empezaron a desarrollar estrategias de evaluación que se vinculaban más estrechamente a los objetivos curriculares (Kulm, 1990; *Mathematical Sciences Education Boarad*, 1993; Morrow y Smith, 1990; Valencia y Calfee, 1991). Por otra parte, la investigación en psicología cognitiva y motivacional aportó tanto la teoría como las evidencias, gracias a las cuales se perfiló el camino para los cambios que se necesitaban (Black, y Wiliam, 1998; Crooks, 1988; Pellegrino, Chudowsky y Glaser, 2001). Por último, este nuevo modelo de evaluación en el aula, se ha hecho manifiesto en un nuevo tipo de libro de texto de

* El concepto del que habla la autora es *accountability*, que se ha traducido como *responsabilización* porque se responsabiliza a escuelas, directores y maestros del progreso académico de los estudiantes. Este concepto, en el contexto de la educación, hace referencia al uso sistemático de datos de evaluación y otro tipo de información para garantizar que las escuelas vayan en la dirección deseada. Con frecuencia, en los sistemas de responsabilización se incluyen metas, indicadores de progreso hacia el cumplimiento de esas metas y análisis de datos, así como procedimientos de información prescritos y consecuencias o sanciones. La responsabilización a menudo incluye el uso de resultados de evaluación y otros datos para determinar la eficacia del programa y para tomar decisiones sobre recursos, recompensas y consecuencias. (Nota de la traductora)[N. T.]

evaluación fundamentado en la práctica docente (Stiggins, 2001; Taylor y Nolen, 2005). Al señalar la naturaleza fundamental de esta transformación, unos cuantos expertos en mediciones han empezado a preguntar cómo deberían cambiar las ideas tradicionales de validez y confiabilidad en el contexto del aula (Brookhart, 2003; Macmillan, 2003; Moss, 2003; Smith, 2003).

En este capítulo se presenta, tanto la concepción como los fundamentos de la investigación sobre las estrategias de evaluación en el aula concebidas para ser parte integral de la enseñanza y el aprendizaje. Comienza con una introducción histórica para explicar el punto de vista que sostuvieron ciertos teóricos de la medición en el pasado sobre la aplicación de pruebas en el aula. Se detiene específicamente en los puntos de vista presentados en ediciones previas de *Educational Measurement*, obras editadas por Lindquist (1951), Thorndike (1971) y Linn (1989), respectivamente. La intención es identificar las ideas que han perdurado, así como las que actualmente se impugnan. La parte principal del capítulo está organizada en tres secciones: 1) Evaluación formativa, 2) Evaluación sumativa y calificación, 3) Evaluaciones externas y en gran escala. En una sección de conclusiones, se consideran las implicaciones de estas ideas transformadoras para el campo de la medición educativa. Se propone un programa de investigación y se sugieren los cambios que se necesitan en la conceptualización de la validez y la confiabilidad para los objetivos del aula.

1. PANORAMA HISTÓRICO DE LA MEDICIÓN EDUCATIVA Y LAS AULAS

El movimiento que pugnaba para que se aplicaran pruebas de rendimiento escolar —que se inició en 1908 con la publicación por Thorndike y sus alumnos de pruebas de Aritmética y Escritura— estaba estrechamente relacionado con el movimiento de la administración científica o movimiento de la eficiencia social. Sus líderes, que consideraban que las escuelas estaban fallando (U. S. Congress, *Office of Technology Assessment*, 1992), elaboraron instrumentos para documentar la necesidad de mejorarlas y establecer el rumbo para lograrlo. Se pedían pruebas estandarizadas para que pudieran agregarse resultados de diversas escuelas y compararlas entre sí; y el *nuevo tipo* de examen objetivo se consideró como un remedio para la *escandalosa* falta de confiabilidad de los exámenes que hacían los maestros, y que había quedado demostrada en varios estudios anteriores (Thorndike, 1922). Desde el principio hubo también críticos que se quejaron de que las pruebas objetivas medían *tan sólo hechos o fragmentos de información* en vez de que midieran *la capacidad de razonar y la aptitud para la organización*, etcétera, (Wood, 1923). No obstante, al hablar en nombre del pensamiento dominante, Wood estableció los argumentos fundamentales de las ventajas de la medición objetiva[1], los cuales fueron repetidos a lo largo de todo el siglo:

> La prueba estandarizada es sumamente exacta no sólo debido a que uno puede calificarla objetivamente, sino porque da suficientes muestras del desempeño del examinado así como del material sobre el cual se examina. Por otra parte, es lo bastante flexible como para que pueda examinar no sólo *información* sino también juicio, y la evaluación de relaciones, causas y consecuencias. (p. 162)
>
> No hay tanta oposición entre *información* y *razonamiento* como algunos maestros quisieran hacernos creer ... Los hechos no son sólo un aspecto legítimo e indudable del pensamiento... sino que sólo el pensamiento puede adquirirlos, conservarlos y reproducirlos, y sólo puede hacerse esto mediante la organización lógica y sistemática del material. (p. 162)
>
> Todo estudio experimental del que tenga-

[1] Para ser justos, Wood (1923) argumentó en favor del uso de pruebas estandarizadas por su mayor confiabilidad con el fin de *complementar* mediciones que utilizaban los exámenes tradicionales de ensayo. Sin embargo, otros usaron posteriormente este mismo razonamiento para sustituir pruebas de ensayo por pruebas objetivas más confiables.

> mos conocimiento, y que se haya realizado hasta ahora, ha mostrado una fuerte relación entre la medición de la información en un campo y la inteligencia o la capacidad de pensar en el material de ese campo. (p. 163)

Por otra parte, puesto que las pruebas de rendimiento se elaboraron en el mismo periodo y las hicieron los mismos autores de las pruebas de Coeficiente Intelectual (CI), ambos tipos de pruebas llegaron a tener los mismos formatos de reactivos y los mismos modelos estadísticos cuyas raíces se hundían en la psicología de las diferencias individuales.

Después de la Primera Guerra Mundial, el uso de pruebas estandarizadas de rendimiento aumentó notablemente debido al éxito práctico y en gran escala de la prueba *Army Alpha*, al establecimiento de grandes oficinas de investigación educativa y agencias de investigación cooperativa, y a la conceptualización de Ralph Tyler en el campo de la evaluación educativa, con el propósito de evaluar cuán bien habían logrado sus objetivos los programas educativos (Cook, 1941; Madaus y Stufflebeam, 2000; U. S. Congress, *Office of Technology Assessment*, 1992).

La primera edición de *Educational Measurement*, libro que se publicó en 1951 y que editó E. F. Lindquist, reflejó y extendió la idea de que las pruebas estandarizadas eran esenciales para el proceso de evaluación y mejoramiento de la educación. Si bien los autores de los capítulos de la edición de 1951 dicen que las "funciones de la medición educativa se relacionan... con la facilitación del aprendizaje" (Cook, 1951, p. 4), lo que ellos tenían en mente era que esto lo realizarían pruebas elaboradas fuera del aula. Su punto de vista era que había que desarrollar programas de pruebas estandarizadas en los distritos escolares, que es lo que hoy podría llamarse administración de datos o sistemas manejados con datos. La distinción que se hace en nuestro informe contemporáneo *Knowing What Students Know* (Pellegrino et al., 2001) —entre los tipos de datos de evaluación que se necesitan para las políticas en gran escala, en contraposición con las decisiones rutinarias que se toman en el aula— no era una distinción evidente en 1951.

Walter Cook (1951), en su capítulo titulado *The Functions of Measurement in the Facilitation of Learning*, defendía el uso de la medición objetiva para adaptar la enseñanza a las necesidades individuales de aprendizaje. En tanto que reconocía que "Los mejores maestros no dejan de seguir constantemente el proceso de verificación del aprendizaje mediante la observación directa de la conducta y las pruebas informales", al mismo tiempo argumentaba en favor del valor de las *pruebas preparadas por expertos* porque:

1. Son más concienzudamente analíticas que la mayoría de las que podrían preparar los maestros.
2. Hacen que los maestros tomen conciencia de los elementos importantes, las secuencias necesarias y las dificultades del proceso.
3. Ahorran tiempo y energía al maestro para hacer diagnósticos, y le dejan más tiempo y energía para que haga el trabajo de corrección.
4. Ayudan al alumno a reconocer sus necesidades de aprendizaje al hacer hincapié en forma sistemática en sus errores.
5. Los procedimientos correctivos se indican o proporcionan al maestro y le ahorran tiempo, así como también le ayudan a sistematizar el proceso. (p. 37)

En el capítulo sobre *The Functions of Measurement in Improving Instruction*, que se relaciona con lo anterior, Tyler (1951) observó que "la medición educativa está concebida, no como un proceso totalmente distinto de la enseñanza, sino más bien como una parte integral de ésta" (p. 47). Pese a la similitud entre estas palabras y las concepciones actuales de la evaluación en el aula, los procesos que Tyler tenía en mente se asumían casi totalmente *fuera* del aula. Si bien algunos de los ejemplos de Tyler tomaban en cuenta la posibilidad de que un instructor individual de un curso pudiera recorrer todo el proceso de planeación de la enseñanza —especificando objetivos, planificando experiencias de aprendizaje y valoran-

do efectos—, Tyler se interesaba sobre todo en que un programa de pruebas de rendimiento en un distrito escolar se "planificara y desarrollara como parte integral del programa del currículo y la enseñanza" (p. 64). Los maestros aprenderían participando en el desarrollo de objetivos y en la elaboración de pruebas, y también aprenderían de los datos resultantes.

Para la segunda edición, de 1971, Robert Glaser, un psicólogo cognitivo, y Anthony Nitro, un teórico de la medición, escribieron el capítulo sobre *Measurement in Learning and Instruction*. Su perspectiva formal de planificación de la enseñanza estaba influida por supuestos docentes conductistas y de la educación adaptativa por computadora, que gozaban de popularidad en la época. Al igual que Tyler y Cook, su visión planteaba la necesidad de pruebas pertinentes para la enseñanza, que se elaborarían fuera del aula y se entregarían a los docentes. "A medida que se desarrolla la enseñanza, la información para tomar decisiones educativas deben darse al docente, al estudiante y posiblemente a una máquina" (pp. 626-627). En vista de la complejidad que implica el diseño de pruebas apropiadas, incluyendo la validación de modelos cuantitativos de adaptación de la enseñanza, "parecería además que el agobio de planificar y elaborar tales pruebas, de procesar las respuestas y de llevar a cabo análisis preliminares de los resultados de la prueba, debe manejarlo alguien que no sea el maestro de la clase." De nueva cuenta, al igual que Tyler, Glaser y Nitko vieron que estos resultados de las pruebas podrían estar al servicio de la evaluación de programas, al proporcionar una retroalimentación formativa a los responsables del sistema. Sin embargo, a diferencia de Tyler, estaban más interesados en datos de las pruebas que pudieran usarse en forma permanente para adaptar la enseñanza a los estudiantes individuales, y creían que su sistema de pruebas diseñado externamente podría quedar incluido en la enseñanza de un modo perfectamente consistente. "Si se hacen apropiadas y sutiles, la docencia, la educación y la aplicación de pruebas se irán diluyendo una en otra". (p. 646)

Al escribir él solo para la tercera edición, Nitko (1989) ofreció nuevamente un enfoque sobre la planificación de la enseñanza que aprovechaba una base de investigación cognitiva sumamente sofisticada con el fin de ofrecer pruebas pertinentes, desde el punto de vista de la instrucción, para que se utilizaran en el aula. Revisó la literatura sobre pruebas de diagnóstico que se basaban en los conocimientos previos requeridos, en habilidades y en el dominio de objetivos conductuales –acercamientos que reflejaban una idea muy estrecha de cómo se desarrolla el aprendizaje y una representación relativamente empobrecida del dominio de los contenidos. En contraste con esto, Nitko (1989) también estudió textos emergentes de diagnóstico centrados en el análisis de errores y estructuras de conocimiento de los estudiantes. Estas últimas categorías le permitieron vislumbrar los temas de investigación que serían fundamentales para la concepción contemporánea del aprendizaje y la evaluación. Por ejemplo, "(la) comprensión que tenga quien diseña pruebas sobre el significado y la estructura del conocimiento que un estudiante trae al sistema de enseñanza" podría utilizarse para identificar "la comprensión cotidiana de palabras y fenómenos [...] que no concuerden con la comprensión canónica de los expertos" (p. 461). Es importante señalar que Nitko (1989) también observó que "la investigación reciente en psicología educativa indica que las formas en que los estudiantes representan mentalmente el conocimiento son tan importantes, para el desarrollo de sus habilidades para resolver problemas y para el aprendizaje avanzado, como las formas en que manifiestan su conocimiento conductualmente" (p.466). Si bien Nitko (1989) todavía imaginaba un sistema de enseñanza informal y de evaluación que se elaboraría fuera del aula y se entregaría a los maestros, su punto de vista reflejó un cambio esencial e importante: quedaron lejos las pruebas de competencia del estudiante después de una lección o periodo de instrucción, donde pasaba o reprobaba, y se dio un paso hacia las evaluaciones más ricas de su comprensión y aprovechamiento en un campo del conocimiento.

Si bien los primeros volúmenes de *Educational Measurement* tenían poco que decir sobre las pruebas que hacían los docentes o sobre las prácticas de evaluación en el aula, el movimiento de pruebas estandarizadas y el paradigma de la evaluación de programas determinaron, no obstante, qué se enseñaba a los docentes acerca de la evaluación. Los teóricos de la medición, responsables de los cursos de *Pruebas y Mediciones* para maestros, creían que debía enseñarse a éstos cómo emular la confección de pruebas estandarizadas de rendimiento, así como de qué manera debían usar una diversidad de mediciones estandarizadas. Los libros de texto típicos desde la década de 1940 hasta la de 1990 incluían los siguientes capítulos:

I. Finalidad de la medición y la evaluación
II. Análisis estadístico de los resultados de pruebas
III. Validez
IV. Confiabilidad
V. Principios generales de la elaboración de pruebas (incluye la especificación de los objetivos de la enseñanza)
VI. Principios de la elaboración de pruebas objetivas
VII. Principios de elaboración de pruebas de ensayo
VIII. Análisis de reactivos para pruebas en el aula
IX. Asignación de calificaciones y forma de reportarlas
X. Pruebas de CI y aptitud académica
XI. Pruebas estandarizadas de rendimiento
XII. Mediciones de intereses y personalidad
XIII. Interpretación de normas estadísticas para las pruebas

Los libros de texto de medición se centraban en la elaboración de pruebas formales cuya finalidad era la asignación de calificaciones. Si bien varios autores mencionaban la importancia de utilizar la información que aportaban las pruebas para modificar la enseñanza, los libros de texto daban pocas explicaciones sobre el modo en que los docentes entenderían los datos con objeto de rediseñar su enseñanza. Algunos autores dieron por sentado que sería sencillo "volver a enseñar ciertas cuestiones" (Torgerson, y Adams, 1954). La estadística y las presentaciones cuantitativas de la confiabilidad y la validez fueron sobresalientes en cuanto a qué necesitaban saber los docentes. En el prefacio a su texto de *Educational Measurement*, Travers (1955) señaló que muchos de sus colegas tenían preferencia por los libros de texto sobre medición *psicológica* porque brindaban "un alimento intelectual más sólido que los libros sobre medición *educativa*" (p. vi), en vista de lo cual aumentó el nivel técnico de su libro "para ayudar a fortalecer una debilidad en la preparación de los maestros" (p. vi). En una muestra de treinta libros de texto examinados para este análisis histórico, sólo encontré dos que tenían una sección o subsección dedicada al uso de la observación en el aula. Encontré un texto que menciona el uso de la evaluación para la retroalimentación[2]. La mayor parte de los textos tenían un capítulo sobre la especificación de objetivos para la enseñanza; y, si se captara el mensaje principal de estos libros, ayudaría a los maestros a volverse más sistemáticos en el uso que hacen de diversos formatos de reactivos para representar un contenido importante. Sin embargo, sería justo decir que los aspectos técnicos de la elaboración de pruebas recibieron más atención que las conexiones entre la evaluación y las actividades de la enseñanza.

El acento que ponen los integrantes de la comunidad de la medición en temas formales y técnicos también puede encontrarse en la literatura de investigación sobre la capacitación en medición para los maestros. Esta literatura, que se extiende por varias décadas, se lamenta constantemente de que los maestros han recibido una preparación deficiente para cumplir con

[2] Brown (1981), de manera profética, comentó: "(El) aspecto eficaz de la retroalimentación es saber si una pregunta se contestó correctamente o, si se contestó en forma incorrecta, saber dónde ocurrió el error y qué necesita hacerse para corregir dicho error. Esta información es suministrada por el señalamiento del maestro sobre la corrección de la respuesta y/o por sus comentarios, no por la calificación. Así, una prueba que se corrige, pero que no da lugar a una calificación, puede proporcionar tanta retroalimentación útil a los estudiantes como una que si lleva a una calificación" (p. 171).

sus responsabilidades de medición y evaluación. Históricamente, muchos estudios dieron por sentado que los maestros necesitaban saber lo que se enseñaba en los cursos de *Pruebas y Mediciones*, y hablaban de la idoneidad de dicha preparación según cuántos programas formativos ofrecían tales cursos, cuántos estados los requerían y cuántos maestros los tomaban (Goslin, 1967; Noll, 1955; Roeder, 1972; Ward, 1980). Cuando los investigadores intentaron identificar en forma directa las habilidades específicas que se juzgaban esenciales para los maestros, los instrumentos que utilizaron en sus encuestas limitaron desafortunadamente la base de conocimientos posibles a los contenidos de los libros de medición. Así, por ejemplo, en el estudio de mayo de 1964, se pidió a maestros y directores así como a profesores de universidad y expertos en pruebas, que clasificaran la importancia de setenta competencias. Treinta y dos de éstas eran elementos estadísticos que tenían que ver con el cálculo y la interpretación de la media, la mediana y la moda, la desviación estándar, las puntuaciones estándar, las correlaciones y así sucesivamente. Las competencias restantes que recibieron clasificaciones más altas, correspondieron fielmente a los capítulos de los libros de texto de medición citados anteriormente. De igual manera, en 1973, Goehring utilizó libros de texto de *Pruebas y Medición* para generar una lista de 116 competencias y luego pidió a maestros y directores de escuelas que analizaran su importancia relativa.

Sólo alguna que otra vez, y relativamente en fecha reciente, han empezado los especialistas de medición a fijarse en el contexto del aula para tratar de entender las necesidades que tienen los maestros de obtener competencia en evaluación. En 1973 Fahr y Griffin cuestionaron la literatura que, al parecer, hacía de la medición un fin en sí mismo, y sostuvieron en cambio que las habilidades deben estar relacionadas directamente con las decisiones de enseñanza que los maestros necesitan tomar. Varios estudios, centrados inicialmente en cómo se utilizaban en el aula las pruebas estandarizadas, revelaron la gran importancia de las pruebas elaboradas por los maestros, las pruebas incluidas en el currículo, y las interacciones y observaciones informales para su toma de decisiones cotidiana (Dorr-Bremme, 1983; Salmon-Cox, 1981; Yeh, Herman y Rudner, 1981). Gracias a datos de entrevistas, Dorr-Bremme (1983) concluyeron que los maestros actúan como razonadores prácticos y como clínicos, y que orientan sus actividades de evaluación a las tareas prácticas que deben de llevar a cabo en las rutinas cotidianas, tales como "decidir qué enseñar y cómo enseñarlo a los estudiantes de diferentes niveles de desempeño; llevar el registro de cómo progresan los alumnos y cómo ellos (los maestros) pueden ajustar su enseñanza apropiadamente y evaluar y calificar a los estudiantes en su desempeño" (p. 3). Para estos fines, los maestros se apoyan en muy buena medida en las pruebas hechas por ellos y en las interacciones con los estudiantes y las observaciones que hacen de éstos. Stiggins y Conklin (1992) publicaron resultados de una serie de estudios de campo que documentaban qué hacen los maestros para evaluar a sus estudiantes, y analizaron qué necesitarían saber los docentes para hacer bien estas tareas. Stiggins (1991) concluyó que la capacitación tradicional en medición ha estado *crónicamente mal enfocada*, tanto así, que sólo nosotros tenemos la culpa por la falta de atención dada a la capacitación en medición en los programas educativos para maestros.

Muy aparte de la literatura sobre medición, los expertos en los contenidos de la enseñanza empezaron a elaborar alternativas a las pruebas estandarizadas para el aula, movidos por una aversión a los efectos de las pruebas de rendición de cuentas o responsabilización, pero también a causa de profundos cambios en las concepciones del aprendizaje y el aprovechamiento en las materias (Shepard, 2000). En lectura, por ejemplo, los investigadores que trabajaban desde una perspectiva emergente del alfabetismo (Clay, 1985; Teale y Sulzby, 1986), se dedicaron mucho más a observar y respaldar las habilidades en desarrollo de niños en contextos sociales concre-

tos, más que en habilidades aisladas y descontextualizadas. Clay (1985) inventó estrategias de evaluación insertas en el acto de leer, señalando que la investigación había fracasado en demostrar que el aprendizaje hubiera mejorado por la asignación de estudiantes a grupos de diferente nivel con base en un diagnóstico asentado en pruebas de velocidad en lectura o en pruebas de prerrequisitos, tales como habilidad lingüística o discriminación visual. Goodman (1985) reintrodujo el concepto de *vigilancia de los niños*, un remanente del movimiento de estudios del niño, que estuvo en boga mucho tiempo antes. A diferencia de las pruebas estandarizadas que se aplican en un único momento, la vigilancia de niños es continua. Legitima la importancia de la observación profesional en el aula y tiene en cuenta experiencias de aprendizaje más ricas que aquellas que pueden *sepultarse sin riesgos en la prueba* (Goodman, 1985, p. 14). La reunión de muestras del trabajo de los estudiantes (Teale, Hiebert y Chittenden, 1987) se convirtió en un recurso valioso, no sólo para alcanzar una comprensión del pensamiento de los niños, sino para documentar su progreso a lo largo del tiempo. Se descubrió que la narración de cuentos era más eficaz, tanto para los fines de la enseñanza como para los de evaluación, que las preguntas tradicionales de comprensión de lectura (Morrow, 1985), y así sucesivamente. Irónicamente, estos investigadores estaban haciendo exactamente lo que Tyler había recomendado décadas antes al clarificar sus objetivos de enseñanza y buscar una representación lo más fiel posible de estos objetivos en sus planes de evaluación. Con todo, lo que Tyler no previó fueron las limitaciones de los formatos de las pruebas objetivas, que para finales del siglo XX ya no eran adecuadas para corresponder a las nuevas concepciones del aprendizaje de las materias (Shepard, 2000).

En la comunidad de las Matemáticas, la fuerza motriz y la dirección para los cambios en la evaluación fueron paralelos a los que hubo en Lectura. En el capítulo sobre Matemáticas del tercer *Handbook of Research on Teaching*, Romberg y Carpenter (1986) observaron que había ocurrido un cambio considerable en la investigación sobre la enseñanza y el aprendizaje. Debido a la revolución de la ciencia cognitiva, la investigación ya no se centraba solamente en conductas observables, sino que también tomaba en consideración procesos cognitivos internos. En vez de un modelo de aprendizaje en el que los maestros transmiten el conocimiento y los estudiantes lo absorben, el nuevo modelo de aprendizaje sostiene que los estudiantes construyen activamente conocimiento nuevo. En las obras del *National Council of Teachers of Mathematics* (Consejo Nacional de Maestros de Matemáticas [NCTM], por sus siglas en inglés) *Curriculum and Evaluation Standards for School Mathematics* (1989) y *Everybody Counts*, un informe del *National Research Council* (1989), el aprendizaje de las Matemáticas se redefinió, más bien, como un proceso de indagación y de búsqueda de sentido que una *repetición y mímica sin sentido*. Para la evaluación esto significaba que se necesitaban problemas más extensos y no rutinarios para atraer a los estudiantes y para evaluar la *potencia matemática* –definida como la capacidad de usar el conocimiento matemático "para razonar y pensar creativamente y para formular, resolver y reflexionar críticamente en los problemas" (NCTM, 1989, p. 205). Además, el discurso del aula se convirtió en un punto focal de las reformas en Matemáticas que buscaba proporcionar a los estudiantes la oportunidad de conjeturar y explicar su razonamiento. Estas nuevas rutinas de enseñanza incrementaron al mismo tiempo la importancia de las evaluaciones informales e integradas -observaciones, preguntas del maestro y escritura de un diario- como medios para comprender el pensamiento de los estudiantes (Silver y Kenney, 1995). En vez de la práctica imperante según la cual los maestros ajustaban sus propias pruebas para que emularan tanto la forma como el contenido de las pruebas externas de opción múltiple, Silver y Kilpatrick (1989) sostuvieron que debía hacerse un esfuerzo serio para *dotar de nuevas habilidades* a los docentes, para que pudieran dar clases con enfoque de resolución de problemas y para evaluar las capacidades y actitudes de sus

estudiantes en cuanto a resolución de problemas, en el contexto de tales clases.

Pueden narrarse historias parecidas para otras materias, como ciencias y estudios sociales. En cada caso, los reformadores de finales del siglo XX estuvieron motivados por la teoría constructivista del aprendizaje y la necesidad de una enseñanza y una evaluación más auténticas (Resnick y Resnick, 1992; Wiggins, 1993). Para la comunidad dedicada a las mediciones en Estados Unidos, el impacto de estos cambios se concentró principalmente en reformar programas de evaluación en gran escala, ya que varios estados iniciaron un programa de evaluaciones innovadoras y basadas en el desempeño (Baron y Wolf, 1996). Quizá a causa de la gran importancia de las pruebas de responsabilización externa[3], la comunidad estudiosa de las mediciones mostró lentitud en considerar las implicaciones de estos cambios teóricos para la evaluación en el aula. Se formó un pequeño *Grupo de Interés Especial* dentro de la *American Educational Research Association*; pero, durante los años noventa, por ejemplo, la evaluación en el aula o temas afines, tales como las formas de asignar calificaciones, dieron cuenta de sólo el 4% de las sesiones en las reuniones anuales del *National Council on Measurement in Education* (Consejo Nacional de la Medición en Educación). En Gran Bretaña, las pruebas estandarizadas tomaron una dirección muy diferente. El Assessment Reform Group (1999), que empezó en 1989 como un Grupo de Tareas [*Task Group*] de la *British Educational Research Association*, se centró en el vínculo decisivo entre la evaluación en el aula y la enseñanza y el aprendizaje. El *Assessment Reform Group* acuñó la expresión *evaluación para el aprendizaje* para referirse a la evaluación que respalda el proceso de aprendizaje, lo que contrasta con la evaluación que sólo mide los resultados del aprendizaje. Siguiendo a Sadler (1989), Black y Wiliam (1998) hicieron que este aprendizaje se centrara en la característica definitoria de la evaluación formativa, diciendo que la evaluación es formativa "sólo cuando la comparación de los niveles reales y los de referencia producen información que luego se usa para modificar la laguna" (p. 53).

En la siguiente sección sobre la evaluación formativa, me explayo sobre la idea del uso de la evaluación como parte del proceso de aprendizaje. La evaluación formativa se define como la evaluación llevada a cabo durante el proceso de enseñanza con el fin de mejorar la enseñanza o el aprendizaje. La evaluación formativa puede implicar métodos informales, tales como la observación y las preguntas orales, o el uso formativo de medidas más formales como exámenes tradicionales, portafolios y evaluaciones del desempeño. También me ocupo de problemas de coherencia y de cómo podríamos lograr que las estrategias de evaluación formativas y sumativas se respaldaran mutuamente. La distinción entre la evaluación formativa y la sumativa es paralela al uso original que Michael Scriven (1967) dio a estos términos, en el contexto de la evaluación curricular y la evaluación de programas, para distinguir entre la evaluación realizada durante el proceso de desarrollo para dar información al proceso mismo, en contraposición con la evaluación del producto final. La *evaluación sumativa*, que se considera en una sección posterior, se refiere a las evaluaciones realizadas al final de una unidad de enseñanza o curso de estudio, con el propósito de dar calificaciones o de certificar el aprovechamiento del estudiante. Como veremos, el nuevo modelo de evaluación formativa aspira a hacer que la evaluación forme parte integral de la enseñanza, tal como lo propusieron los primero teóricos de la medición. La diferencia importante es que las estrategias que se explican aquí están construidas sobre un modelo muy diferente de enseñanza y aprendizaje, y no dependen de instrumentos estandarizados que se hayan elaborado fuera del salón de clase.

[3] *External accountability*. Hace referencia a que la sociedad responsabiliza y, también pide una rendición de cuentas a la institución, a los maestros y/o a los estudiantes mismos. La interna sería la propia de la institución, los maestros y/o los estudiantes [N. T.]

2. EVALUACIÓN FORMATIVA

Para que los docentes sean eficaces en reforzar el aprendizaje de los estudiantes, deben comprobar constantemente la comprensión que éstos vayan logrando. Por otra parte, tienen que darles a conocer la importancia de que ellos mismos asuman la responsabilidad de reflexionar y supervisar su propio progreso en el aprendizaje. Un análisis fundamental de Black y Wiliam (1998), que marcó un hito, descubrió que los esfuerzos orientados a mejorar la evaluación formativa producían beneficios mayores a la mitad de una desviación estándar. En otras palabras, la evaluación formativa, eficazmente implementada, puede hacer tanto o más para mejorar la realización y los logros que cualquiera de las intervenciones más poderosas de la enseñanza, como la enseñanza intensiva de Lectura, las clases particulares y otras parecidas.

En esta sección, comienzo con un resumen de las teorías contemporáneas del aprendizaje y luego presento un modelo de evaluación formativa del que muestro su compatibilidad tanto con la teoría cognitiva como con la teoría sociocultural del aprendizaje. Luego considero varias estrategias y herramientas específicas que comprenden el modelo general que los docentes utilizan como parte de las rutinas cotidianas de enseñanza. Estos procesos recursivos de evaluación son esenciales para una revisión y perfeccionamiento continuos de la enseñanza así como para mejorar también el aprendizaje del estudiante.

No obstante, antes es necesario hacer una advertencia. Las prácticas ideales de evaluación que aquí se explican y que se basan en la investigación, son consistentes con las prácticas de maestros particularmente competentes y expertos, pero no necesariamente reflejan prácticas de evaluación típicas. De hecho, la mayoría de los maestros en servicio tiene solo un conocimiento limitado de estrategias de evaluación formativa, y sigue pensando en la evaluación como un proceso que sirve principalmente para calificar. Por consiguiente, la sección final sobre la investigación futura debe considerar el aprendizaje y el desarrollo profesional de los maestros así como a la eficacia de herramientas específicas de evaluación.

2.1 Teoría del aprendizaje, y coherencia en el diseño de la evaluación

La obra Knowing What Students Know (Pellegrino et al., 2001) fue el resultado de un comité del *National Research Council*, que se encargó de reunir los avances hechos tanto en la ciencia cognitiva como en la medición. Una premisa central que sustenta las recomendaciones de *Knowing What Students Know* es que las observaciones e interpretaciones de la evaluación deben estar relacionadas con un modelo cognitivo bien estructurado de cómo aprende el estudiante en cierto campo. Este modelo fundamental debe reflejar una comprensión actualizada de cómo se desarrolla el aprendizaje en un campo, y no las "creencias tan restrictivas" (Pellegrino et al., 2001, p. 54) en que se basan las evaluaciones del logro académico que más se utilizan. Un modelo de aprendizaje sirve "como un elemento unificador, un núcleo que da cohesión al currículo, la enseñanza y la evaluación" (Pellegrino et al., 2001, p. 54). Por otra parte, los autores de *Knowing What Students Know* argumentan más adelante en favor de esta misma *coherencia* esencial entre las evaluaciones externas y las que se hacen en clase. Para trabajar conjuntamente y respaldar el aprendizaje del estudiante, las evaluaciones en ambos niveles de un sistema de evaluación deben apoyarse en modelos compatibles sobre el aprendizaje del estudiante, aun cuando los modelos referentes al aula pueden ser mucho más detallados. En este capítulo, utilizo el concepto de coherencia para hablar acerca de cómo puede hacerse que se respalden mutuamente la evaluación formativa y la sumativa dentro del aula.

En el panorama histórico que esbocé en líneas anteriores, los cambios en la teoría del aprendizaje se mencionaron varias veces como la fuerza motriz de los cambios en la forma en que conceptualizaron la enseñanza y la evaluación

los expertos en las materias. Las concepciones contemporáneas del aprendizaje han transformado nuestra comprensión de cómo ocurre el aprendizaje, pero de manera más fundamental, han alterado nuestras concepciones de qué es el aprendizaje, qué significa ser competente en un campo y, por consiguiente, cómo buscaríamos evidencia de esa competencia. La revolución cognitiva fue una rebelión en contra de la psicología de las diferencias individuales y el conductismo, que se había interesado sobre todo en la adquisición de competencias mediante el reforzamiento de conductas observadas, en lugar de tratar de explicar los procesos mentales básicos. En cambio, de acuerdo con la teoría cognitiva, quienes aprenden construyen el conocimiento conectando nueva información a estructuras previas de conocimiento. Los esquemas mentales, en el cerebro o la mente, sirven para organizar el conocimiento para que cuando después se necesite lo recuperemos y utilicemos en situaciones problemáticas; y los procesos ejecutivos, llamados *metacognición*, permitan monitorear y manejar su propia comprensión y aprendizaje a quienes aprenden. Los cognitivistas hacen hincapié en la comprensión conceptual y han demostrado que la transferencia, esto es el uso del conocimiento en situaciones nuevas, se hace posible por la aprehensión de principios generalizados y el uso de esquemas gracias a los cuales se reconocen las similitudes en los distintos tipos de problemas. La perspectiva cognitiva predomina en el trabajo teórico y empírico presentado en dos publicaciones significativas del *National Research Council*, *How People Learn* (Bransford, Brown, y Cocking, 1999) y *Knowing What Students Know*, si bien ambas reconocen la influencia del contexto social y la aplicabilidad de las perspectivas socioculturales o situacionales.

Un modelo sociocultural del aprendizaje surge de un renovado interés en el trabajo de Vygotsky (1978) y otros psicólogos rusos. Se ocupan de la naturaleza social del aprendizaje y en la idea de que la competencia y la identidad de quienes aprenden se desarrollan por medio de una participación mediada socialmente, en una actividad práctica llena de significado. Un individuo aprende a pensar y a razonar gracias a una gran diversidad de apoyos proporcionados por adultos y pares más conocedores e informados. Este modelo de aprendizaje queda excelentemente caracterizado como un proceso de inducción, o como un modelo del aprendiz, mediante el cual se permite a los novatos participar y hacer contribuciones en un contexto de trabajo real, pero a los que se asignan tareas adaptadas a su nivel particular de competencia. El ejemplo paradigmático, desde luego, es la forma como se adquiere el lenguaje gracias a un proceso de práctica mediada socialmente (Bruner, 1985). En la teoría sociocultural es fundamental comprender que los productos de la actividad (los resultados del aprendizaje) están integrados en las prácticas culturales del ambiente donde se desarrolla la actividad. De este modo, aprender a conocer es volverse un adepto que participa en las maneras de hablar, las representaciones del conocimiento y el uso de herramientas asociadas con una comunidad de práctica específica.

En otro lado he propuesto una concepción del aprendizaje *social-constructivista*[4], que reúne las teorías cognitiva y sociocultural (Shepard, 2000). Si bien hay muchos puntos discutibles no resueltos entre estas dos perspectivas y en el interior de ellas, las considero compatibles. Vygotsky (1978, p. 57) sostenía que "todas las funciones en el desarrollo cultural del niño aparecen dos veces: primero, en el nivel social, y luego en el nivel individual; primero entre la gente (*interpsicológica*), y luego *dentro* del niño (*intrapsicológica*)." Siguiendo la dirección de Vygotsky, pues, podemos usar la investigación sociocultural para entender los procesos sociales que respaldan y definen el aprendizaje, y la teoría cognitiva, con el fin de entender los procesos mentales subsecuentes y recurrentes del individuo. Obsérvese, sin embargo, que incluso el razonamiento y las reflexiones de una persona aparentemente pri-

[4] En este capítulo no defino ni uso el término social-constructivismo como la conjunción de las teorías cognitiva y sociocultural, porque algunos autores utilizan el social-constructivismo en forma más estrecha para referirse solamente a una variante de la teoría cognitiva.

vadas, están integradas socialmente, porque el individuo lleva consigo las formas de razonar, las expectativas, los criterios, etcétera. de su mundo social. Más recientemente, en mi trabajo me he concentrado en la teoría sociocultural porque creo que es la teoría más completa.

2.2 Un modelo de la evaluación formativa

Sadler (1989) aportó el modelo más aceptado de la evaluación formativa. Este autor indicó que es insuficiente que los maestros simplemente den una retroalimentación respecto de si las respuestas son correctas o incorrectas. En vez de ello, para facilitar el aprendizaje, es igualmente importante que la retroalimentación esté vinculada explícitamente a criterios claros de desempeño y que se proporcione a los estudiantes estrategias de mejoramiento. Este modelo de evaluación formativa fue explicado más ampliamente en un reporte reciente de Atkin, Black, y Coffey (2001) sobre evaluación en ciencias en el aula. Estos autores construyen el proceso de evaluación del aprendizaje con estas preguntas clave:

- ¿Adónde tratas de ir?
- ¿Dónde estás ahora?
- ¿Cómo puedes llegar ahí?

Al responder la pregunta de evaluación (la No. 2, ¿dónde estás ahora?) en relación con el objetivo de la enseñanza (pregunta No. 1) y dedicándose específicamente a lo que se necesita para alcanzar el objetivo (pregunta No.3), el proceso de evaluación formativa respalda directamente el mejoramiento.

Establecer objetivos claros para el aprendizaje por parte del estudiante implica mucho más que anunciar una finalidad de la enseñanza para que los estudiantes la contemplen. También requiere la elaboración de los criterios mediante los cuales será juzgado el trabajo del estudiante. ¿Cómo sabrán el maestro y el estudiante que se ha entendido un concepto? ¿Cómo se evaluará la capacidad del estudiante para defender un argumento? Luego, la fase de evaluación debe ocurrir durante el proceso de aprendizaje, mientras el estudiante trabaja en tareas que ejemplifican directamente el objetivo del aprendizaje que se propone alcanzar. Esta evaluación, que se hace en medio del aprendizaje, podría ocurrir por medio de preguntas al estudiante durante el trabajo grupal, cuando un/una estudiante explica a la clase cómo resolvió un problema, o al examinar un trabajo escrito. Finalmente, en la tercera etapa, para que la evaluación formativa sea de verdadera ayuda para el aprendizaje, debe darse una retroalimentación que proporcione entendimiento acerca de cómo llenar una carencia. Por ejemplo, cuando un estudiante tiene aún confusión respecto de un concepto fundamental, ¿existe un método diferente para abordar el problema o un conocimiento esencial al que haya que volver? Si el razonamiento en un trabajo de composición está desarrollado en forma deficiente, ¿cómo puede corregirlo el estudiante después de tomar primero en consideración lo que falta en relación con los criterios de evaluación?

Este modelo de evaluación formativa es más que una etapa de recolección de de datos. Es un modelo para el aprendizaje que corresponde directamente a la Zona de Desarrollo Próximo (ZDP) y a la teoría sociocultural del aprendizaje. Tal como lo visualizó Vygotsky (1978), la Zona de Desarrollo Próximo es la región, en un continuo imaginario de aprendizaje, entre lo que un niño puede hacer de manera independiente y lo que ese mismo niño puede hacer si lo ayudan. Wood, Bruner y Ross (1976) desarrollaron además la idea del *andamiaje*[5] para caracterizar el apoyo, bajo la modalidad de guía, indicaciones y estímulo, que los adultos proporcionan en la ZDP con objeto de capacitar a quien aprende a desempeñar a un nivel de logro algo que de otro modo no habría sido capaz de alcanzar. La etapa de evaluación en el modelo de evaluación formativa (¿dónde estás ahora?) proporciona la

[5] *Scaffolding* en inglés. Se refiere a la técnica mediante la cual el maestro modela la tarea o la estrategia deseada del aprendizaje, y luego traslada gradualmente la responsabilidad a los estudiantes.

comprensión que se necesita para un respaldo eficaz. Y el modelo formativo completo, que comprende el esclarecimiento del objetivo y la identificación de los medios para llegar ahí, puede verse esencialmente como un sinónimo del andamiaje de la enseñanza. En realidad, la versión plenamente elaborada de Sadler de la evaluación formativa requiere que los maestros y los estudiantes tengan una comprensión y una apropiación compartidas del objetivo de aprendizaje y, por último, que los estudiantes sean capaces de supervisar su propio mejoramiento. Esto corresponde al objetivo del andamiaje: fomentar que quien aprende interiorice este proceso y asuma su responsabilidad.

En el mundo real, los maestros rara vez tienen tiempo para dar sesiones de tutoría individuales o de hacer evaluaciones dinámicas que les permitieran dedicarse a impartir una enseñanza con andamiaje a un estudiante durante un ciclo completo de aprendizaje. Y, con seguridad, planear la enseñanza para un salón de clases completo lleno de estudiantes cuya ZDP es muy variada, es un verdadero reto. No obstante, pueden establecerse rutinas del aula para garantizar que los elementos básicos de la evaluación formativa y del andamiaje estén establecidas y funcionando en la forma de interacciones de enseñanza ordinarias. Por ejemplo, hablar con estudiantes en forma individual es una parte normal de la enseñanza de la escritura, junto con la corrección editorial por los compañeros o el ejercicio de *la silla del autor*, que es una práctica en la que un estudiante comparte algo que ha escrito con toda la clase. Otra estrategia es desarrollar las capacidades de los estudiantes para proporcionarse retroalimentación unos a otros. Una de las razones por las que el discurso sobre el aula ha recibido tanta atención en la investigación sobre la reforma en la enseñanza, es que las pautas de interacción grupal, especialmente las preguntas de los estudiantes y la forma en que explican su razonamiento, pueden proveer de andamiaje al aprendizaje del estudiante, sin que sea necesario que el docente invierta tiempo para atender a los estudiantes uno por uno. Cobb, Wood, y Yackel (1993) hablan de análisis de toda una clase donde hay andamiaje y donde los estudiantes son capaces no sólo de esclarecer su comprensión de conceptos matemáticos, sino también de practicar las normas y formas sociales de hablar en esa disciplina. De igual manera, en clases de investigación en Ciencias (Hogan y Pressley, 1997), los estudiantes aprenden a dar evidencias que sustenten una posición y también a criticar las conclusiones sin fundamento de sus compañeros de clase: una forma valiosa de retroalimentación. Semejantes despliegues públicos en que se desarrolla el pensamiento crean, asimismo, la oportunidad perfecta para la evaluación formativa.

Idealmente, pues, la evaluación formativa debe quedar perfectamente integrada en la enseñanza. En los párrafos que siguen me explayo sobre los elementos específicos del proceso de evaluación formativa, los cuales tienen un extenso fundamento de investigación. Comienzo ocupándome del contenido porque la evaluación carece de significado si no abarca las cosas que más queremos que aprendan los estudiantes. En seguida considero los avances de los estudiantes porque, dentro de los contenidos de las materias, los maestros deben tener también una idea de las progresiones típicas de aprendizaje, con el fin de que sepan en qué dirección ayudan a los estudiantes, y también cómo respaldarlos cuando la comprensión falla. Más tarde tomo en cuenta aspectos específicos de las interacciones entre evaluación y enseñanza, cómo tener acceso al conocimiento previo, cómo hacer explícitos los criterios, proporcionar retroalimentación, etcétera. No obstante, no hace falta que ninguno de estos procesos interrumpa la enseñanza, sino que más bien deben retroalimentar un aprendizaje continuo. Aun si uno toma tiempo para un examen formal, los resultados pueden utilizarse para un diagnóstico de la enseñanza con el fin de decidir qué conceptos necesitan más análisis y trabajo. Y los estudiantes pueden llegar a entender que tales evaluaciones tienen una finalidad de aprendizaje.

2.3. La importancia del contenido: selección de las tareas de enseñanza y de evaluación que encarnan objetivos de aprendizaje

La evaluación no puede impulsar el aprendizaje si se basa en tareas o preguntas que distraen la atención de los verdaderos objetivos de la enseñanza. Históricamente, las pruebas tradicionales a menudo dirigían erradamente la enseñanza cuando se centraban en lo que era más fácil de medir en vez de centrarse en lo que era importante aprender. La enseñanza en clase debe ocupar a los estudiantes en actividades de aprendizaje, las cuales sean lo más directamente posible ejemplos de los objetivos reales del aprendizaje. Si queremos que los estudiantes sean capaces de leer libros, periódicos y poemas, ellos deben en realidad hacer eso, por lo que no hay que darles materiales abreviados y simulados, excepto cuando queramos adaptarlos a su edad. De igual modo, en Ciencias, si queremos que los estudiantes sean capaces de razonar y usar el conocimiento científico, entonces debemos darles la oportunidad de explicarse cómo funcionan las cosas, realizando investigaciones y elaborando explicaciones con sus propias palabras, para que así conecten sus experiencias con las teorías del libro de texto. La evaluación, entonces, debe realizarse como parte de estas actividades de aprendizaje significativo. Si los estudiantes realizan un proyecto de investigación en Historia o muestran a la clase cómo resolvieron un problema de Matemáticas, entonces la tarea de la enseñanza es la labor de la evaluación.

Un rasgo definitorio de las reformas basadas en estándares ha sido el desarrollo de normas curriculares que sirven para revigorizar y realzar qué significa saber y demostrar aprovechamiento en cada disciplina. Por ejemplo, el *Curriculum and Evaluation Standards for School Mathematics* del NCTM (1989)[6] plantea expectativas, hace hincapié en la solución de problemas, la comunicación, el razonamiento matemático y el establecimiento de conexiones que van mucho más allá del dominio de las habilidades y los conceptos básicos. No es de sorprender que la reforma de la evaluación fuera una parte igualmente importante de los movimientos de estándares, por la necesidad de alcanzar estos objetivos más ambiciosos. El término *alineación* se ha utilizado para especificar la correspondencia deseada entre evaluaciones y estándares curriculares. Desafortunadamente, el significado de alineación se desvaloriza un tanto cuando los editores de pruebas muestran que todos su reactivos de selección múltiple pueden hacerse corresponder con las categorías de los estándares de contenido de un estado, aun cuando en conjunto sólo utilizan un estrecho subconjunto de los estándares deseados. Anteriormente, he propuesto el término *encarnar* como una forma de caracterizar mejor la alineación más completa y real que ocurre cuando las tareas, los problemas y los proyectos en los que están ocupados los estudiantes representan toda la gama y profundidad de lo que decimos que queremos entiendan y tengan capacidad de hacer (Shepard, 2003).

Tal como lo ilustraron Wiggins y McTighe (1998), la elaboración de evaluaciones que manifiesten objetivos de aprendizaje es esencial para una buena enseñanza, y no sólo una cuestión de medir resultados. En vez de un planeamiento de la enseñanza que se concentre en actividades interesantes, Wiggins y McTighe utilizan un proceso de *planificación hacia atrás*, que empieza con los objetivos de la enseñanza, luego se cuestiona ¿cuál será una evidencia convincente o una demostración de que hubo aprendizaje? y al final planea actividades que permitirían desarrollar en los estudiantes esa comprensión. Con esta última como objetivo de la enseñanza, subrayar la evaluación fuerza a los maestros a explicar en forma muy clara y sin lugar a dudas qué aspecto tendría la evidencia de la comprensión, y estas descripciones del desempeño los mueven a brindar oportunidades a los estudiantes para que desarrollen y practiquen estas habilidades que de otro modo podrían haberse perdido si la *comprensión* se hubiese dejado sólo como un objetivo declarado para toda la unidad. Por ejemplo, un

[6] Siglas en inglés del *National Council of Teachers of Mathematics*, Consejo Nacional de Maestros de Matemáticas. [N. T.]

maestro encontraría pruebas de comprensión por parte de los estudiantes si éstos pudieran explicar su razonamiento o aplicar su conocimiento en un contexto nuevo. Y cada una de estas facetas puede desarrollarse aún más para dejar en claro qué es lo que los estudiantes serían capaces de hacer. Por ejemplo, el criterio de Wiggins y McTighe (1998) para una explicación que demuestre comprensión incluye dar razones creíbles, proporcionar una explicación sistemática, o utilizar modelos mentales útiles. Si bien es cierto que ser capaz de explicar un concepto exige mucho más que conocerlo, conocer y explicar guardan una estrecha relación entre sí, y el razonamiento y pensamiento que además se requieren para producir una explicación creíble son exactamente el tipo de esfuerzo mental que hace falta para desarrollar una comprensión más flexible y más profunda.

2.4. Progresiones del aprendizaje

Las progresiones del aprendizaje o los continuos del aprendizaje son importantes para monitorear y respaldar el desarrollo a lo largo del tiempo. A diferencia de los estándares, que han recibido una gran atención durante la última década, se ha desarrollado relativamente poco y ha habido mucha menos investigación para explicar las progresiones en el aprendizaje. Desde luego que la mayoría de los maestros tienen cierto sentido intuitivo de qué sigue, o no serían capaces de ayudar a los estudiantes a desempeñarse mejor. Sin embargo, incluso los maestros más capacitados podrían sacar provecho de modelos con mayor desarrollo formal sobre cómo se despliega el aprendizaje en un ámbito curricular, y se beneficiarían también si conocieran las variaciones y desviaciones naturales del patrón típico. Si bien las progresiones empíricamente validadas pueden hacer posible un andamiaje más comprensivo de la enseñanza, las progresiones jamás deben interpretarse como algo cerrado o como una secuencia absoluta de requisitos.

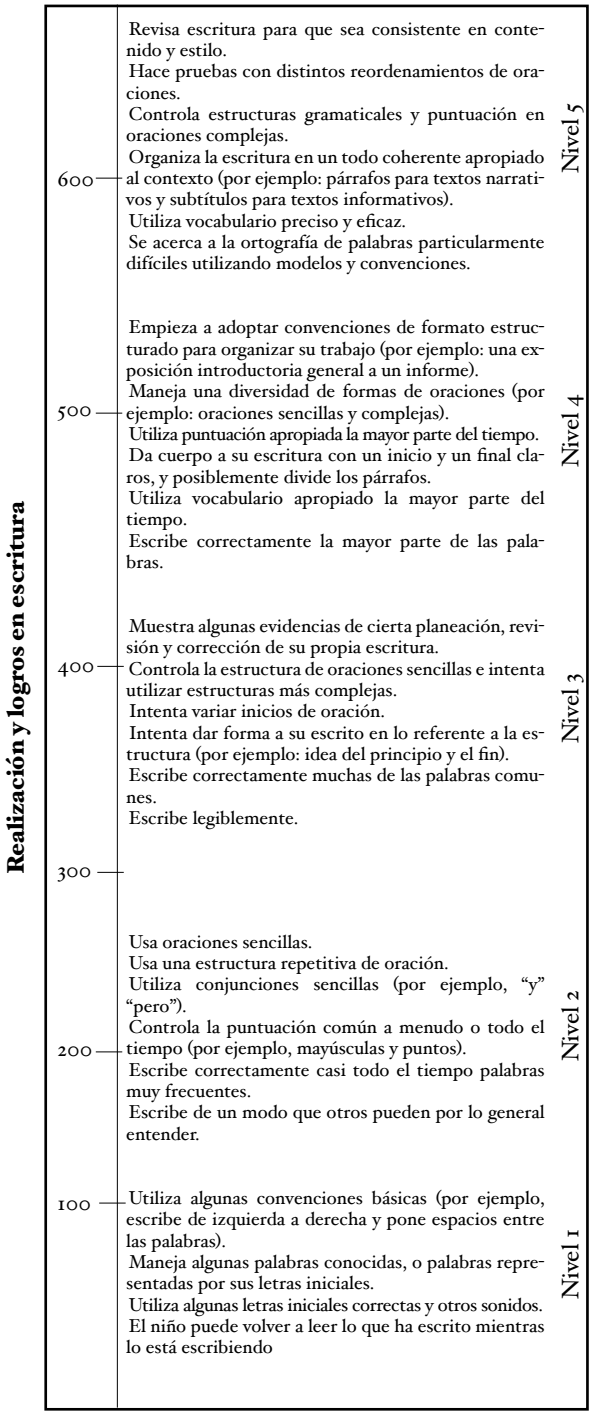
El *mapa de progreso* en escritura que aparece en la Figura 17.1, del *Australia's National School English Literacy Survey,* el cual ejemplifica una progresión en el aprendizaje, fue diseñado para reunir datos básicos sobre la realización y los logros de estudiantes del tercer y quinto grados en el curso de Lengua, incluyendo Habilidades de lectura, Revisión visual, Audición y Expresión oral y escrita (Masters y Forster, 1997). El progreso de un estudiante en conseguir el control de estructuras y convenciones del lenguaje puede diagramarse o comunicarse en este *continuum*, el cual proporciona una imagen del crecimiento individual en contraposición con un telón de fondo de expectativas establecidas normativamente. A diferencia de los informes de evaluación, que se ven más bien como una lista de verificación de objetivos de los diferentes grados, los mapas de progreso tienen implicaciones más directas para la enseñanza, porque brindan simultáneamente una imagen de las fortalezas y debilidades y una manera de mirar hacia lo que viene para cada una de las facetas del campo de que se trate. Por ejemplo, un estudiante de segundo grado puede estar muy avanzado en el uso de la puntuación (mayúsculas y puntos) y deletrear palabras comunes correctamente (Nivel 3), pero tal vez necesite ayuda para experimentar e ir más allá de estructuras repetitivas de oraciones (Nivel 2).

En Australia, los sistemas de evaluación son coherentes con las evaluaciones en el aula y en gran escala están relativamente bien desarrollados y vinculados a un mismo mapa de progreso (Forster y Masters, 2004). Algo parecido sucede en los Países Bajos, donde se elaboran *trayectorias aprendizaje-enseñanza* para proporcionar los conocimientos pedagógicos necesarios para respaldar el desarrollo del pensamiento del estudiante a lo largo del tiempo (Van der Heuvel-Panhuizen, 2001). En Estados Unidos, en cambio, el desarrollo de progresiones de aprendizaje útiles para la enseñanza se ha visto limitado por la forma transversal y a retazos, que con el tiempo han propiciado los sistemas de evaluación en gran escala. Las evaluaciones estatales y nacionales, que originalmente se propusieron monitorear tendencias en momentos particulares,

centraron la atención en las expectativas del nivel a alcanzar en ciertos grados fundamentales (cuarto año de primaria, segundo de secundaria y tercero de preparatoria). En años más recientes, los estados, que ahora piden más requisitos para las pruebas individuales, completan los grados intermedios e intercalan las expectativas curriculares. No obstante, estas expectativas, especialmente en momentos en que los *estándares de nivel internacional* se han establecido en currículos que nunca se habían implementado antes, no necesariamente reflejan la trayectoria del desarrollo de estudiantes reales. Al mismo tiempo, existe el peligro de que depender de promedios normativos para establecer progresiones, congelará las expectativas curriculares pasadas de moda o establecerá expectativas demasiado bajas –porque *promedian* los resultados de una enseñanza fallida. Por ejemplo, con base en datos empíricos, la *Keymath Diagnostic Arithmetic Test* (Connolly, Nachtman, y Pritchett, 1972) utiliza reactivos en la forma de 4 1/2 x 5 como ejemplo de la Aritmética que, se espera, deben dominar los estudiantes en el grado equivalente al tercero de secundaria.

Lo que hace falta es un proceso para crear continuos de aprendizaje que estén basados no sólo en la investigación, sino en un juicio experto que incluya la validación de los continuos propuestos en el contexto de un currículo bien implementado. Hasta la fecha, algunas de las investigaciones más fundamentadas y más pertinentes respecto de la enseñanza, se han hecho en las áreas del alfabetismo emergente y de la habilidad con los números también emergente, las cuales pueden servir como modelo. Por ejemplo, en la ortografía prefonémica, los niños escriben primero letras que representan palabras sin relacionar los sonidos de las letras con la palabra deseada. Después dan un paso en su desarrollo cuando empiezan a escribir letras que tienen correspondencia con los fonemas más prominentes en una palabra, y así sucesivamente (Hiebert y Raphael, 1998). De modo parecido, las estrategias naturales de los niños para sumar se desarrollan con el tiempo, desde que empie-

FIGURA 17.1 Mapa de progreso en escritura fundamentado empíricamente



Fuente: De Masters, G. y Forster, M. (1997), *Mapping Literacy Achievement: Results of the 1996 National School English Literacy Survey*, Canberra, Australia, Department of Employment, Education, Training and Youth Affairs (DEETYA). Reproducido con permiso.

zan a contar hasta que usan números (Carpenter y Moser, 1984), y la investigación sobre la enseñanza ha demostrado que los maestros se vuelven más eficaces cuando se les hace tomar conciencia de estas estrategias típicas de solución de problemas (Fennema y Franke, 1992).

2.5. Evaluación del conocimiento previo

El conocimiento previo es esencial para el aprendizaje. De hecho, el proceso del aprendizaje puede concebirse como lo que hacemos para conectar e integrar una nueva comprensión con el conocimiento existente. El conocimiento previo incluye el aprendizaje formal, como el de un preescolar que aprende la norma de no cruzar la calle sin mirar hacia los dos lados, pero también incluye una multitud de explicaciones implícitas, las cuales nos enseñamos a nosotros mismos, sobre cómo funciona el mundo. Estas intuiciones o teorías que nos enseñamos a nosotros mismos pueden en ocasiones facilitar nuevo aprendizaje, como cuando las explicaciones científicas se dominan fácilmente porque *tienen sentido* y armonizan con nuestra experiencia previa. Las teorías intuitivas también pueden ser el origen de conceptos erróneos importantes que obstaculizan el aprendizaje nuevo y son relativamente impenetrables al cambio que busca la enseñanza, a menos que los estudiantes reciban una forma estructurada de trabajo que les permita resolver las inconsistencias entre sus intuiciones y otras evidencias.

Las estrategias eficaces de enseñanza se basan en el conocimiento previo de los estudiantes como recurso. Por otra parte, al usarse rutinas de activación del conocimiento al principio de nuevas lecciones y unidades de estudio, los maestros ayudan a los estudiantes a desarrollar el hábito de preguntarse (cuando se enfrentan a un nuevo aprendizaje o a una tarea en la que tengan que resolver un problema) *de lo que ya sé, ¿qué puede ayudarme a resolver esto?* Muchas actividades de conocimiento previo -como las conversaciones instruccionales (Tharp y Gallimore, 1988) y las técnicas K-W-L[7] (Ogle, 1986)- no son consideradas como evaluaciones como tales ni por maestros ni por estudiantes. Sin embargo, sí aportan datos valiosos para corregir la enseñanza, como cuando los maestros encuentran lagunas en un conocimiento que suponen ya asimilado o cuando descubren que los estudiantes saben mucho más sobre un tema de lo previsto. Dada la evidencia obtenida, gracias a la investigación, sobre la necesidad de combatir conceptos erróneos cuando se presenten, el reconocimiento explícito de los conceptos erróneos evaluados como la razón para las actividades posteriores de enseñanza podría ser una manera de elevar la conciencia de los estudiantes de que la evaluación está al servicio de la finalidad del aprendizaje.

El conocimiento previo es más que un conjunto de hechos que un estudiante ha acumulado en su casa y en grados anteriores. El conocimiento previo también incluye patrones de lenguaje y formas de pensar que los estudiantes desarrollan por medio de sus roles sociales y sus experiencias culturales. Los maestros pueden a veces interpretar mal las diferencias en las prácticas culturales y considerarlas como evidencia de un *déficit*. Por ejemplo, los niños blancos de clase media están más acostumbrados a que les hagan preguntas descontextualizadas, como *¿qué color es éste*?, que los niños de otros grupos sociales (Heath, 1983). Las reglas implícitas de interacción pueden hacer que a los maestros se les dificulte advertir las fortalezas de los estudiantes fuera de su propio grupo social, a menos que dispongan de medios para sacar esas fortalezas de una manera que tenga sentido culturalmente. Por ejemplo, Moll, Amanti, Neff y González (1992) utilizan el concepto *fondos de conocimiento*

[7] *Instructional conversations* son charlas en las que se exploran ideas más que respuestas a preguntas de exámenes que se evaluarán. Técnicas KWL:
K Representa la ayuda que se da a los estudiantes para que recuerden lo que *Saben [know*, en inglés] sobre la materia.
W – Representa la ayuda que se da a los estudiantes para que determinen lo que *Quieren* [want, en inglés] aprender.
L - Representa la ayuda que se da a los estudiantes para que identifiquen lo que *Aprenden* [*learn*, en inglés] al leer. [N. T.]

para referirse al conocimiento que reciben en casa los niños de familias pobres, y que se basa en la agricultura, la carpintería, la medicina, la religión, el cuidado infantil y las actividades relacionadas con el manejo del presupuesto, que pueden usarse para reforzar el conocimiento escolar.

2.6. Criterios explícitos y el uso de guías de calificación (rúbricas)

El modelo de evaluación formativa requiere que el maestro y el estudiante tengan una comprensión compartida de los objetivos del aprendizaje. En la teoría cognitiva, las metas deben definirse explícitamente y ser visibles para los estudiantes. En la teoría sociocultural, una comprensión del objetivo se construye conjuntamente mientras el estudiante recibe ayuda para mejorar su desempeño. Cuando los maestros ayudan a los estudiantes a entender e internalizar los estándares de excelencia de una disciplina —es decir, aquello que hace que un trabajo de historia o una explicación matemática sean buenos— les ayudan a desarrollar la conciencia metacognitiva de aquello a lo que necesitan prestar atención mientras escriben o resuelven un problema. Ciertamente, aprender las normas y formas de una disciplina es parte del aprendizaje de ésta, y no sólo un medio para sistematizar o justificar la calificación. Por otra parte, no es probable que los estudiantes lleguen a entender qué significan los estándares de excelencia simplemente porque el maestro coloque en la pared las guías de calificación [*scoring rubrics*], aunque éstas pueden ser un punto de referencia útil. Más bien, los estudiantes desarrollan la comprensión de las expectativas por medio de la retroalimentación y de las autoevaluaciones, gracias a las cuales los criterios se vinculan directamente a sus propios esfuerzos de aprendizaje.

2.7. Retroalimentación

Uno de los hallazgos más antiguos de la investigación psicológica (Thorndike, 1931) es que la retroalimentación facilita el aprendizaje. Sin retroalimentación —sobre errores conceptuales o retrocesos ineficaces— es probable que el que aprende persista en cometer los mismos errores. En un extenso meta-análisis de 131 estudios controlados, Kluger y DeNisi (1996) dieron a conocer un tamaño de efecto [*effect size*] o ganancia de 0.4 gracias a la retroalimentación. También reconocieron una variación significativa en el estudio con aproximadamente un tercio de los estudios que mostraron efectos negativos. Al intentar identificar las características de retroalimentación más asociadas con los efectos positivos, Kluger y DeNisi descubrieron que es más probable estimular el aprendizaje cuando la retroalimentación se enfoca en ciertos aspectos de la tarea y destaca los objetivos de aprendizaje. Este importante hallazgo proveniente de la literatura sobre la retroalimentación, es consistente con mi argumento anterior para las guías de calificación [*rubrics*], las cuales permiten juzgar el desempeño en relación con criterios bien definidos (en vez de juzgar a un estudiante comparándolo con otros), y se armoniza con los descubrimientos de la literatura sobre motivaciones que analizaré posteriormente en el contexto de las prácticas de calificación.

De acuerdo con evidencias de la investigación, es un error hacer elogios falsos, tratando de motivar a los estudiantes y aumentar su autoestima. Al mismo tiempo, la retroalimentación negativa directa, sin consideraciones, puede minar el aprendizaje y la disposición del estudiante a esforzarse más. Por consiguiente, una comprensión de las consecuencias motivacionales de la retroalimentación es tan importante como conocer sus propósitos cognitivos. El modelo de evaluación formativa, consistente con la literatura cognitiva, demuestra que la retroalimentación es especialmente eficaz cuando dirige su atención a cualidades particulares del trabajo del estudiante en relación con criterios establecidos y proporciona una guía sobre qué hacer para mejorar. Además, los maestros deben establecer un clima de confianza y desarrollar normas en clase que posibiliten la crítica cons-

tructiva. Esto significa estratégicamente que la retroalimentación debe ocurrir durante el proceso de aprendizaje (y no al final, cuando ya se terminó el aprendizaje de ese tema); maestro y alumnos deben tener una comprensión compartida de que la finalidad de la retroalimentación es facilitar el aprendizaje; y puede significar que la calificación debe quedar en suspenso durante la etapa formativa.

Para que haya una retroalimentación eficaz, es necesario que los maestros sean capaces de analizar el trabajo del estudiante e identificar los patrones de errores y las lagunas que más atención requieren (no cualquier error posible). En un estudio de intervención, Elawar y Corno (1985) descubrieron que los maestros mejoraban extraordinariamente la eficacia de la retroalimentación cuando se concentraban en estas preguntas: "¿Cuál es el error principal? ¿Cuál es la razón probable de que el estudiante cometiera este error? ¿Cómo puedo guiar al estudiante para que evite el error en un futuro?" (p. 166). Los maestros también deben entender la teoría de cómo la retroalimentación incrementa el aprendizaje para que ellos puedan desarrollar rutinas en el aula que comprueben la comprensión de los estudiantes y aseguren que no los dejan solos para que persistan en los malos hábitos o los conceptos erróneos.

2.8. Enseñar y evaluar para que haya transferencia

La *transferencia* se refiere a la capacidad de utilizar nuestro conocimiento en contextos nuevos. La transferencia es obviamente una meta del aprendizaje. ¿De qué sirve el conocimiento si no podemos acceder a él o no podemos aplicarlo? Sin embargo, los estudios de la capacidad de los estudiantes para utilizar información pertinente incluso de una lección reciente en la que hubo aciertos son notoriamente decepcionantes. La transferencia se ve inhibida cuando los estudiantes aprenden de memoria y se someten a rutinas mecánicas para resolver problemas sin pensar. Los ejemplos son numerosos, y van desde reagrupar (o pedir prestado) de memoria en aritmética de segundo grado a las soluciones a problemas del tipo *inserta y a ver si pega* en los cursos de física en la universidad[8]. En cambio, la investigación sobre expertos y novatos y los estudios de transferencia nos demuestran que es más probable que haya transferencia cuando el aprendizaje inicial se centra en la comprensión de principios fundamentales, cuando se consideran explícitamente las relaciones de causa y efecto y sus razones, y cuando los principios de aplicación están presentes en forma directa.

Dar clases para que haya transferencia requiere que la enseñanza inicial se centre en la comprensión. También significa trabajar ostensiblemente para ampliar la comprensión de los estudiantes. Por ejemplo, debe ser algo común —tan pronto como se vea que los estudiantes han dominado un nuevo tipo de problema o una manera de resolver un problema— que los maestros hagan una nueva pregunta relacionada con ese conocimiento pero lo amplíe. El problema inicial y las aplicaciones de seguimiento en la Figura 17.2 provienen de *Connected Mathematics* (Lappan, Fey, Fitzgerald, Friel y Phillips, 1998). Éstas son tareas ejemplares en varios sentidos. En primer lugar, ilustran que las buenas tareas de evaluación pueden ser intercambiables con las buenas tareas de enseñanza. Segundo, la investigación completa de la que se han extraído estos problemas ayuda claramente a desarrollar un entendimiento del principio fundamental de equivalencia, al pedir a los estudiantes que establezcan relaciones entre representaciones tabulares, gráficas y algebraicas. Tercero, los problemas de aplicación pueden pensarse como tareas de cuasi-transferencia que pueden utilizarse

[8] *Plug and chug* es la expresión original en inglés: En los cursos introductorios de física en la universidad, los estudiantes empiezan a resolver un problema recurriendo a la solución algebraica y numérica, es decir buscan ecuaciones y las manipulan, insertando números en las ecuaciones hasta encontrar una combinación que produzca una respuesta. Rara vez utilizan su conocimiento conceptual de física para analizar cualitativamente la situación del problema, ni tampoco planean sistemáticamente una solución antes de que empiecen con sus manipulaciones numéricas y algebraicas de las ecuaciones. Cuando llegan a una respuesta, generalmente se muestran satisfechos y rara vez verifican si la respuesta tiene sentido. [N. T.]

para estar seguros de que los estudiantes pueden generalizar lo que han aprendido en la investigación. Sin embargo, los estudiantes tendrán que pensar un poco sobre las características únicas de la nueva tarea. No pueden simplemente aplicar la regla del Problema 1 de memoria a las Aplicaciones 1 y 2.

Una de las razones por la cual los expertos tienen mejores habilidades de transferencia que los novatos, es porque son capaces de reconocer peculiaridades de los problemas que son las mismas y diferentes a las de problemas resueltos con anterioridad. Por consiguiente, es importante que los estudiantes aprendan a pensar específicamente sobre cómo pueden utilizar lo que ya saben. En este sentido, enseñar para manejar estrategias de transferencia, especialmente de transferencia lejana, también tiene correspondencia con las técnicas de conocimiento previo.

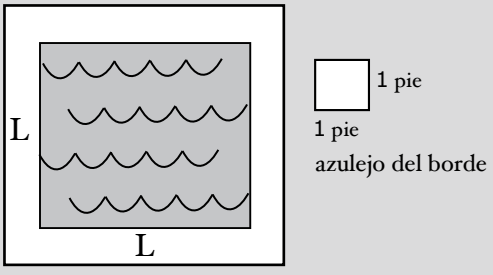
2.9. Auto-evaluación del estudiante

Hacer que los estudiantes se ocupen en criticar su propio trabajo es útil tanto desde el punto de vista cognitivo como desde el motivacional. En esencia el hábito de auto-evaluarse lleva a la auto-supervisión del desempeño, que es la finalidad del andamiaje de la enseñanza así como el objetivo del modelo de evaluación formativa de Sadler (1989). El proceso de auto-evaluación se basa en las ventajas metacognitivas de los criterios explícitos, pues se pide a los estudiantes que piensen y apliquen criterios en el contexto de su propio trabajo. Al hacerlo así, los estudiantes se explican y llegan a entender qué significan los criterios de un modo más profundo que si sólo leyeran una lista de ellos. En términos más generales, este tipo de práctica asistida de metacognición —es decir, la práctica en la cual los estudiantes aprenden estrategias para supervisar su propio aprendizaje— ayuda a desarrollar las capacidades metacognitivas de los estudiantes. Al mismo tiempo, la auto-crítica puede incrementar la responsabilidad del estudiante ante su propio aprendizaje y hacer que la relación entre el maestro y él sea de más colaboración. Esto no significa que los maestros renuncien a su responsabilidad, sino que al compartirla, consiguen que el estudiante tenga mayor posesión, menos desconfianza y más reconocimiento de que las expectativas no son caprichosas ni que están fuera de su alcance.

FIGURA 17.2 Una tarea inicial de evaluación y de enseñanza y tareas de aplicación de transferencia cercana

Problema 1

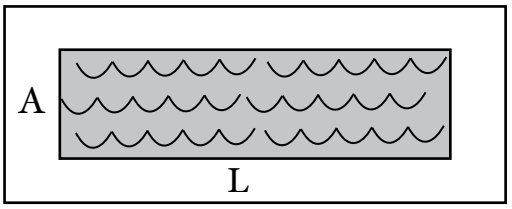
En este problema, explorarás esta pregunta: Si una piscina cuadrada tiene lados de *L* pies de longitud, ¿cuántos azulejos se necesitan para formar el borde?



A. Dibuja en un papel cuadriculado para que te formes una idea de cuántos azulejos se necesitan para los bordes de piscinas cuadradas con lados de longitud 1, 2, 3, 4, 6 y 10 pies. Registra tus resultados en una tabla.
B. Escribe una ecuación para el número de azulejos (*N*) que hacen falta para formar un borde de una piscina cuadrada con lados de longitud de (*L*) pies.
C. Trata de escribir por lo menos una ecuación más para la cantidad de azulejos que hacen falta para el borde de la piscina. ¿Cómo podrías convencer a alguien de que tus expresiones para la cantidad de azulejos son equivalentes?

Aplicaciones

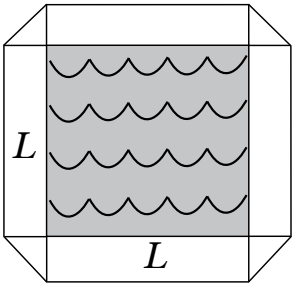
1 **a.** ¿Cuántos azulejos cuadrados de 1 pie se necesitan para formar el borde de una piscina que mide 10 pies de largo y 5 de ancho?
b. Anota una expresión para la cantidad de azulejos del borde que se necesitan para una piscina que mide *L* pies de largo y *A* pies de ancho.



c. Anota una expresión diferente para la cantidad de azulejos que se necesitan. Explica por qué tus expresiones son equivalentes.

2 Una tina cuadrada tiene lados de longitud de *L* pies. Se crea un borde al colocar azulejos cuadrados que miden 1 pie de cada lado a lo largo de los bordes de la tina y azulejos triangulares en las esquinas. Los azulejos triangulares se hicieron cortando azulejos cuadrados a la mitad

a. Si la tina tiene lados de una longitud de 7 pies, ¿Cuántos azulejos cuadrados se necesitan para hacer el borde?
b. Escribe dos ecuaciones para la cantidad de azulejos cuadrados (*C*) que hacen falta para hacer este tipo de borde para una tina cuadrada con lados de longitud de *L* pies.



Fuente: De *Connected Mathematics, Say It with Symbols: Algebraic Reasoning* © 1998 Universidad del Estado de Michigan, Glenda Lappan, James T. Fey, William M. Fitzgerald, Susan N. Friel, y Elizabeth Difanis Phillips, publicado por Pearson Education Inc., publicado como Pearson Prentice Hall. Se ha usado con permiso.

En estudios de caso de dos localidades australianas e inglesas, Klenowski (1995) descubrió que los estudiantes que participaron en una auto-evaluación se interesaron más en los criterios y en la retroalimentación sustantiva que en su calificación misma. Los estudiantes también manifestaron que debían ser más honestos respecto de su propio trabajo, así como también tendrían que ser honestos con otros estudiantes y estar preparados para defender sus opiniones tomando en cuenta la evidencia. Los datos de Klenowski (1995) respaldan la aseveración previa de Wiggins (1992) de que hacer que los estudiantes analicen su propio trabajo hace que se apropien del proceso de evaluación y "hace posible llevar a los estudiantes a niveles más altos porque los criterios son claros y razonables" (p. 30). En un estudio experimental de White y Frederiksen (2000), los estudiantes aprendieron a usar los criterios de la investigación científica para evaluar su propio trabajo. Como parte del protocolo, los estudiantes del grupo experimental tuvieron que escribir una exposición razonada cada vez que se auto-evaluaban, señalando los rasgos distintivos de su trabajo que respaldaban sus valoraciones. Además, los estudiantes utilizaron los criterios para dar retroalimentación a sus compañeros de clase cuando se presentaron proyectos en clase en forma oral. En comparación con el grupo control, los estudiantes que habían participado en la auto-evaluación presentaron proyectos que fueron valorados mucho más por sus maestros (con base en los criterios compartidos). Por otro lado, estudiantes que al inicio tenían realización y logros deficientes mostraron mejoras notables en una medición de comprensión conceptual.

Como parte de las reformas curriculares, los expertos en contenido de diversas materias han desarrollado reformas en la evaluación para integrarlas mejor en la evaluación y la enseñanza. Algunas de estas estrategias, en particular, sirven para hacer que las autoevaluaciones y las evaluaciones por pares formen parte normal de la enseñanza en el aula. Por ejemplo, la técnica de la *silla del autor* es una práctica de enseñanza de lectoescritura en la que los estudiantes aprenden explícitamente las normas para escuchar y dar retroalimentación a sus compañeros de clase respecto de un escrito (Routman, 2000). Hablar con los estudiantes puede también ser un medio para ver si están desarrollando la capacidad de auto-evaluarse. Tanto el estudio de Klenowski como el de White y Frederiksen, ya mencionados, significaron un paso en la auto-evaluación, la cual llegó a formar parte de la enseñanza normal. Es importante mencionar que la finalidad de involucrar a los estudiantes en autoevaluaciones no es llegar a una calificación, sino que ellos obtengan una mayor comprensión, la cual puede utilizarse para nuevos aprendizajes.

2.10 Evaluación de la docencia

El modelo de evaluación formativa se centra en el aprendizaje del estudiante. Un uso igualmente importante de la evaluación en el aula es la evaluación y el perfeccionamiento de la docencia. A la par que los maestros reúnen evidencia acerca de la comprensión del estudiante, también toman en consideración cuáles prácticas docentes funcionan y cuáles no, y qué nuevas estrategias hacen falta. El *Assessment Standards for School Mathematics* del NCTM (1995) identificó tres tipos de decisiones docentes, las cuales fueron informadas a través de datos de evaluación: las decisiones de momento a momento, la planeación a corto plazo y la planeación a largo plazo. Cuando la evaluación y la enseñanza están eficazmente entrelazadas, entonces las ideas de evaluación pueden usarse en tiempo real para ajustar la enseñanza. Por ejemplo, si varios estudiantes cometen el mismo tipo de error, puede ser útil detenerse y dedicar un tiempo al concepto erróneo que subyace a éste. Mientras que la evaluación formativa se centra en qué puede hacer el estudiante para mejorar, la evaluación paralela de la docencia pregunta si los estudiantes han tenido una oportunidad adecuada para aprender.

Los maestros que reflexionan sobre su práctica utilizan datos en forma sistemática para hacer juicios sobre los aspectos específicos de las estrategias docentes que quizá estén obstaculizando el aprendizaje. Buscan explicaciones del éxito o el fracaso en el aprendizaje, y se fijan especialmente en las decisiones de su enseñanza que pudieran ser la causa. Por ejemplo, ¿hay ciertas tareas que parecen hacer que los estudiantes piensen mucho, porque son muy interesantes y dan lugar a múltiples soluciones? ¿Hay algunas actividades que interesan a la mayoría de los niños pero que dejan a las niñas *clavadas en sus asientos*? ¿Batallan con las tareas escolares los chicos que aprenden dos lenguas, si no hay tiempo suficiente para hablar de su conocimiento pertinente de datos esenciales o para esclarecer expectativas? En una revisión ya clásica de su propia enseñanza, Mazur (1997) descubrió que los estudiantes podían resolver problemas como el número 2 de la Figura 17.3, pero no problemas como el número 1. Su extenso análisis de por qué los estudiantes podrían resolver problemas algorítmicos pero no conceptuales y qué hacía él que estimulaba la búsqueda de recetas (incluyendo la forma de sus exámenes) llevó a Mazur a reexaminar su manera de enseñar para enfocarse en estrategias de aprendizaje más activas.

Cuando los maestros utilizan datos de evaluación para modificar su enseñanza, también dan un ejemplo importante a los estudiantes. Tal como sostuve con anterioridad, "si queremos desarrollar una comunidad de estudiosos —en la que los estudiantes busquen en forma natural retroalimentación y critiquen su propio trabajo— entonces es razonable que los maestros modelen el mismo compromiso de usar datos en forma sistemática, ya que esto se aplica a su propio papel en el proceso de enseñanza y aprendizaje". (Shepard, 2000, p. 12)

3. EVALUACIÓN SUMATIVA Y CALIFICACIÓN

La *evaluación sumativa* y *la calificación* constituyen una seria amenaza para los objetivos de aprendizaje declarados por la evaluación formativa. De acuerdo con descubrimientos de la literatura motivacional y de estudios de maestros y estudiantes, las prácticas de las calificaciones pueden minar el proceso de aprendizaje de varias maneras. En primer lugar, las pruebas y las tareas calificadas comunican lo que es importante aprender. Si estas mediciones divergen de las metas del aprendizaje que se valora, entonces los estudiantes concentran su atención y esfuerzo sólo en la porción calificada del currículo. Segundo, el uso de calificaciones como premio o como castigo puede socavar la motivación intrínseca de aprender. Tercero, a aquellos estudiantes para quienes los criterios de las calificaciones les parecen fuera de su alcance, éstas pueden reducir su esfuerzo y su ulterior aprendizaje. Cuarto, la naturaleza comparativa de las prácticas tradicionales de calificación puede reducir la buena voluntad de los estudiantes de ayudar a otros o de aprender de los demás.

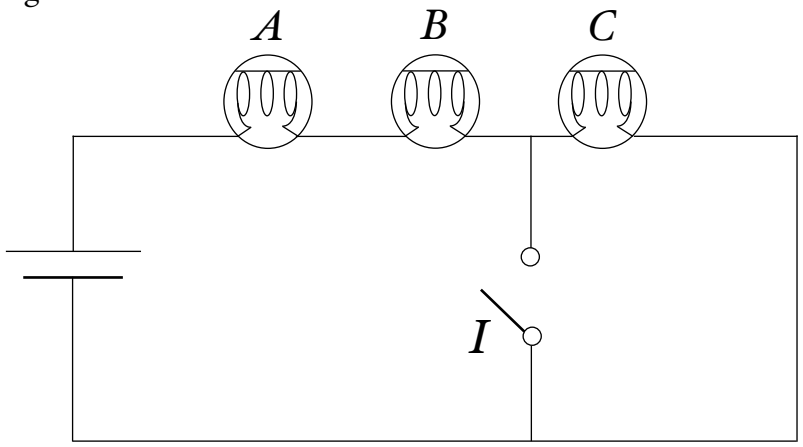
En esta sección, considero qué finalidades se persiguen al calificar y resumo la investigación que se ha realizado sobre las prácticas actuales. Luego, después de examinar los hallazgos empíricos pertinentes de la literatura sobre la psicología cognitiva de la medición y sobre la psicología motivacional, hago un bosquejo de las prácticas de calificación de las que se espera sean no sólo válidas para comunicar logros sino que conduzcan al aprendizaje del estudiante. Si están construidas sobre el mismo modelo fundamental de desarrollar competencia en un campo del conocimiento, entonces es posible que las prácticas formativas y sumativas de evaluación alcancen coherencia y que se respalden mutuamente.

3.1. Finalidades de las calificaciones apropiadas a la edad

Los libros de texto de medición y de psicología educativa dan por sentado que calificar es algo que los maestros deben hacer. Son pocas las razones que se ofrecen, excepto que las políticas escolares exigen que los maestros den calificaciones, a menudo con especificaciones detalladas en cuanto a la información requerida en las boletas. De igual manera, tengo la certeza de que los maestros deben dar calificaciones principalmente porque los padres las desean. Sin embargo, indico que no ha habido estudios sistemáticos de los efectos de las prácticas de calificación sobre la realización y los logros del estudiante. Además, sé que existen tres públicos importantes para las calificaciones: los padres, los usuarios externos, tales como empleadores y funcionarios de servicios de admisión a las universidades, y los estudiantes mismos. Los estudiantes se convierten en el público principal de las calificaciones, porque lo que se les dirá a los demás acerca de sus logros, llega a desempeñar un papel muy importante en las interacciones del aprendizaje. No obstante, si los estudiantes fueran el único público, no queda claro que las calificaciones por sí mismas añadirían información útil. Más bien, lo que maestros y estudiantes necesitan más son evaluaciones sumativas que sirvan para verificar la consecución de logros importantes en la adquisición de competencias por parte de los estudiantes, y las cuales se relacionen con los mismos continuos de desempeño que se utilicen en la evaluación formativa.

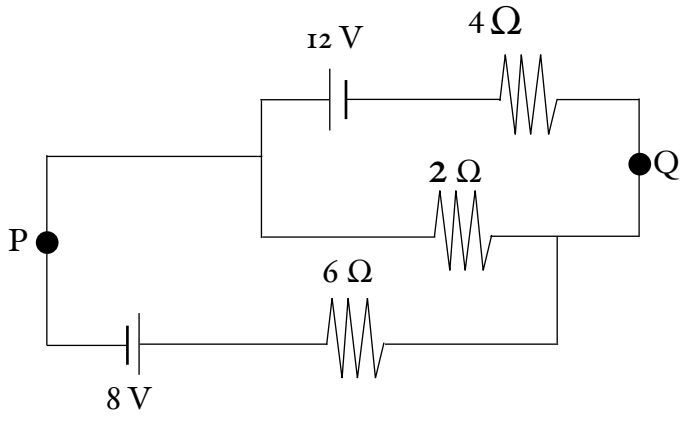
FIGURA 17.3 Pregunta conceptual (arriba) y convencional (abajo) sobre el tema de circuitos eléctricos

1. Un circuito en serie consta de tres focos de luz idénticos conectados a una batería como se muestra aquí. Cuando el interruptor *I* está cerrado, ¿aumentan, disminuyen o permanecen iguales los siguientes?



(a) Las intensidades de los focos *A* y *B*
(b) La intensidad del foco *C*
(c) La corriente que se extrae de la batería
(d) La caída de voltaje entre un foco y otro
(d) La potencia disipada en el circuito

5. Para el circuito que se muestra, calcula *(a)* la corriente en la resistencia de 2-Ω y *(b)* la diferencia potencial entre los puntos *P* y Q.



Fuente: De Mazur, Eric, *Peer instruction: a user's manual*, 1ª edición, © 1997. Reimpreso con permiso de Pearson education, Inc., Upper Saddle River, NJ.

Los principios sobre lo apropiado de las calificaciones en función de la edad, indican que las calificaciones deberían ser mucho menos importantes en la vida del aula en la escuela primaria, de lo que son para los estudiantes de secundaria y preparatoria. Cuando los estudiantes llegan a la secundaria, hay una expectativa mayor de que las calificaciones tengan significado para un público externo. Por ejemplo, una calificación alta en Lengua y Literatura debería implicar que el estudiante puede escribir una composición bien organizada; y un promedio de calificaciones de nueve debe significar que un alumno del último año de preparatoria está bien preparado para el trabajo universitario. Una finalidad básica del movimiento de estándares ha sido, en realidad, la de asegurar este tipo de comprensión compartida sobre el significado de las calificaciones. En cambio, es mucho más difícil llevar a cabo evaluaciones formales de niños pequeños y mucho menos necesario. Los principios articulados por la *National Association for the Education of Young Children* [Asociación Nacional para la Educación de Niños Pequeños, NAEYC por sus siglas en inglés] hacen hincapié en que las evaluaciones de niños pequeños deben basarse en la observación realizada durante las actividades ordinarias del aula, y deben usarse con fines formativos y para comunicárselas a los padres. La NAEYC concluyó que para los niños en los primeros grados "El método de presentar informes a los padres no (debe) depender de las calificaciones con letras o números, sino que más bien (debe) proporcionar una información más significativa y descriptiva de manera narrativa" (p. 15).

La evidencia respecto de qué tipo de información les parece útil a los padres de familia es limitada. En uno de los estudios, sorprendentemente, los padres de algunos niños de tercer año consideraron que lo más valioso era hablar con el maestro sobre el progreso de su hijo; el 77% dijo que esto les parecía muy útil (Shepard y Bliem, 1995). Lo siguiente más útil fue *ver muestras calificadas del trabajo de mi hijo*. El 60% de los padres refirieron que esto fue muy beneficioso en contraposición con las boletas y pruebas estandarizadas, de las que sólo el 43% y el 14% de ellos, respectivamente, consideraron que eran muy útiles. En datos de entrevistas, los padres explicaron que hablar con el maestro les pareció lo más valioso porque les daba un conocimiento de primera mano de las fortalezas y debilidades específicas de su hijo en el contexto del currículo del aula. Una generalización adicional de esta investigación tiene implicaciones significativas para las prácticas de calificación. Los padres quieren comparaciones normativas. No obstante, parecen estar deseosos de tener maestros que les digan cómo va su hijo en lo que respecta a las expectativas que debe responder según el grado, más que exigir una prueba referida a una norma. Esta sustitución de estándares de nivel de grado por comparaciones referidas a una norma, es importante en el contexto de reportes basados en estándares y, como veremos, a causa de las consecuencias desmotivadoras de las prácticas de calificación referidas a las normas estadísticas.

3.2. La investigación en la práctica actual

Brookhart (1994) identificó 19 estudios publicados en los diez años anteriores al trabajo que realizó sobre las prácticas docentes de calificación. Los métodos de estudio fueron variables y abarcaron desde exámenes a maestros hasta estudios de caso. Con el respaldo de resultados muy consistentes entre los estudios, Brookhart (1994) identificó las siguientes generalizaciones:

- Los maestros tratan afanosamente de ser justos con los estudiantes, y en esto se incluye, informarles sobre cuáles serán los componentes de una calificación.
- Las mediciones de rendimiento o desempeño, especialmente las pruebas, son los componentes más importantes para las calificaciones, pero el esfuerzo y la capacidad también suelen tomarse en consideración.
- En las prácticas de calificación hay un efecto del nivel de grado. Los maestros de primaria utilizan evidencia y observación más informales. En el nivel Secundaria,

las mediciones de rendimiento con pruebas de lápiz y papel y otras actividades escritas constituyen una porción mucho mayor de la calificación.

- Hay una variación individual entre las prácticas de calificación por parte de los docentes. Diferentes maestros perciben el significado y la finalidad de las calificaciones en forma diferente, y consideran de manera diferente los factores de desempeño positivo o negativo.

Las prácticas identificadas en los hallazgos 1 y 3 —que los maestros tratan de ser justos y comunican sus criterios de calificación y que usan evidencia más informal en los grados de niveles inferiores— son consistentes con los estándares profesionales basados en la investigación. Sin embargo, el uso de factores de esfuerzo y capacidad para ajustar las calificaciones de desempeño es contrario a las recomendaciones de los expertos en medición y de los reformadores que se basan en los estándares. En un estudio de 143 maestros de escuelas primarias y secundarias del Medio Oeste, Cizek, Fitzgerald y Rachor (1995/1996) descubrieron de modo parecido que el 52% dijo que tomaban en cuenta la capacidad del estudiante individual y el 42% dijo que consideraban el esfuerzo del estudiante al asignar calificaciones. Cizek *et al*. observaron que los maestros se sirven de un conjunto variado de factores distintos al desempeño en formas que crean un *sesgo de buenos resultados*, lo que les ayuda a elevar las calificaciones de los estudiantes. Los expertos dan argumentos en contra de considerar el esfuerzo, la capacidad y la actitud cuando se califica, porque minan la validez de las calificaciones como indicadores de desempeño. Además, estos factores no pueden medirse con precisión, crean desigualdades, invitan a los estudiantes a fingir y confunden a la mayoría o a todos los públicos acerca del significado de las calificaciones.

Al tratar de explicar la gran brecha entre la teoría y la práctica, Brookhart (1994) presenta varios argumentos en nombre de los maestros: el esfuerzo se considera como una parte de *merecer* una calificación; los hábitos de trabajo están estrechamente relacionados con la retroalimentación que los estudiantes necesitan respecto de cómo mejorar, y la participación es esencial para la relación del estudiante con el maestro como instructor. Si bien es decisivo entender los puntos de vista de los maestros y las realidades prácticas del aula, las creencias intuitivas de los maestros acerca de qué es justo y qué motiva a los estudiantes tal vez no cuente con el respaldo de la investigación. También debemos tomar en consideración si los maestros utilizan calificaciones de esfuerzo para *controlar* la conducta de los estudiantes, que no es lo mismo que crear un ambiente de aprendizaje que los *motive*. En un estudio de caso de los métodos de evaluación de los maestros (Lorsbach, Tobin, Briscoe y Lamaster, 1992), un hallazgo de gran importancia fue el que "las tareas y los sistemas de totalizar notas con frecuencia recompensan el haber completado las tareas y la motivación para aprender, y no tanto lo que se sabe" (p. 310). Aparentemente los maestros se desempeñaban dentro de la metáfora de la *escuela es trabajo* (Marshall, 1988), creando un elaborado sistema para estar informados del trabajo del estudiante, pero sin evaluar la calidad o el contenido de ese trabajo.

3.3. Importancia del contenido y del formato: qué se valora

El contenido de las pruebas -qué se evalúa y cómo se evalúa- y el contenido de las tareas que se evalúan para un grado, comunican los objetivos de la enseñanza a los estudiantes y hacen que centren su atención y esfuerzo (obsérvese, sin embargo, que tan sólo dar puntos a tareas no garantiza que pongan atención, si la calidad del trabajo nunca se examina; véase Lorsbach et al., 1992). Anteriormente abordé la cuestión del contenido de la evaluación, en el contexto de la evaluación formativa, y volví a revisarlo cuando examiné el efecto de las pruebas externas en la enseñanza y el aprendizaje en el aula.

Históricamente, los maestros, especialmente en el nivel secundario, se han atenido a las

pruebas formales con vistas a la calificación, y han usado predominantemente en sus pruebas preguntas de bajo nivel. Fleming and Chambers (1983) analizaron 8 mil 800 preguntas de pruebas desde la primaria hasta la secundaria y descubrieron que casi el 80% de ellas tenían el nivel del *conocimiento* en la taxonomía de Bloom (1956). Se descubrieron resultados parecidos una década después. En un estudio nacionalmente representativo, Madaus, West, Harmon, Lomax y Viator (1992) encontraron que el 53% de los maestros de matemáticas en secundaria y preparatoria y el 73% en primaria manifestaron que usaban pruebas del libro de texto al menos una vez al mes. Cuando los analizaron Madaus et al. (1992), sólo 3% de los reactivos en las pruebas finales contenidas en el texto examinaban conocimiento conceptual de alto nivel y alrededor del 5% examinaban habilidades de pensamiento de alto nivel. El restante 95% de los reactivos examinaban información, cálculos y uso de algoritmos y fórmulas en problemas de rutina, parecidos a los que esos estudiantes habían trabajado en el texto. Un porcentaje más alto de maestros comunicaron que hacían sus propias pruebas, pero cuando éstas fueron examinadas en estudios de campo, Madaus *et al*. (1992) descubrieron que eran adaptaciones que se parecían mucho a las pruebas del libro de texto. Más recientemente, Cizek *et al*. (1995/1996) dieron a conocer un importante hallazgo: era más probable que los maestros novatos elaboraran sus propias evaluaciones que los experimentados, quienes tenían la tendencia a atenerse a las pruebas preparadas comercialmente. Este hallazgo podía obedecer al hecho de que los maestros que empiezan tienen más conocimiento de los materiales basados en las reformas y los tipos de problemas y mayor acceso a los mismos.

Como ya he subrayado con anterioridad, la naturaleza del logro en cada una de las disciplinas no puede ser representada adecuadamente por preguntas de bajo nivel en las que sólo se recuerda la información. La reforma de la evaluación ha sido parte integral de la reforma educativa a causa de la necesidad de involucrar a los estudiantes en tareas auténticas con el fin de desarrollar, usar y extender su conocimiento. Un trabajo más significativo dirigido a la comprensión conceptual no sólo proporciona mejores datos de evaluación acerca de cuán bien se desempeñan los estudiantes, sino que también tiene beneficios cognitivos y motivacionales. Por ejemplo, Crooks (1988) revisó estudios que examinaban el vínculo entre el formato de evaluación y las estrategias para estudiar de los estudiantes universitarios. Numerosos estudios corroboraron los primeros hallazgos de Marton y Saljo (1976) de que la aproximación de los estudiantes a las tareas de aprendizaje podría categorizarse como aproximaciones *profundas* o *superficiales*. Las profundas implicaban una búsqueda activa de significado, principios fundamentales y estructuras que relacionaran diferentes conceptos. Las aproximaciones superficiales se concentraban principalmente en la memorización de hechos aislados sin buscar conexiones entre estos hechos. Si bien hay otros factores que influyen en la tendencia de los estudiantes a usar aproximaciones profundas o superficiales, especialmente interés o motivación, en todos los estudios, su percepción de lo que exigían las evaluaciones que esperaban, tuvieron una evidente influencia en la elección de su estrategia.

Los expertos en asignaturas valoran conceptualmente las tareas docentes ricas porque captan lo más importante para que los estudiantes aprendan. Los cognitivistas prefieren tareas estimulantes porque hacen que los estudiantes razonen y favorzcan la generalización si las tareas de transferencia se usan como parte normal de la enseñanza. Las tareas auténticas que requieren un pensamiento de más alto nivel y una activa solución de problemas también incrementan la motivación del estudiante porque son intrínsecamente más interesantes que la memorización o la aplicación de procedimientos sencillos (Stipek, 1998). Por ejemplo, Mitchell (1993) descubrió que las creencias de los estudiantes acerca de la importancia en el mundo real de lo que aprenden era algo que predecía poderosamente su interés y placer en la clase de Matemáticas.

Además, las tareas estimulantes aumentan la motivación intrínseca al intensificar el sentimiento de competencia en los estudiantes. Newmann (1992), por ejemplo, descubrió que los estudiantes califican con los puntajes más altos de interés a las clases que los hacen pensar mucho y los hacen participar activamente en el pensamiento y el aprendizaje.

3.4. La investigación sobre medición, psicología cognitiva y psicología motivacional

A la mayoría de los maestros les disgusta evaluar a sus estudiantes y darles una calificación (Brookhart, 1993; Nava y Loyd, 1992). En vista de los efectos de distorsión de las prácticas de calificación examinados en líneas anteriores, y de los efectos motivacionales negativos de las calificaciones que he considerado, sería fácil ver las calificaciones tan sólo como un mandato desencaminado y opresivo. Sin embargo, hay cierta evidencia de beneficios cognitivos positivos de las evaluaciones sumativas que hay que tomar en consideración junto con los hallazgos de la literatura motivacional. Lo que es más importante, los estudiantes parecen estudiar y aprender más si esperan que les hagan una prueba. Como lo resumió Crooks (1988), las ventajas de poner pruebas pueden explicarse por tres factores. Primero, hacer pruebas de seguimiento hace que los estudiantes revisen y vuelvan a aprender, lo que opera como una forma limitada de práctica distribuida. Segundo, la experiencia misma de presentar pruebas pone a los estudiantes a procesar mentalmente el contenido, si bien esto depende mucho de la calidad de las preguntas en la prueba. Tercero, como ya he observado antes, para bien o para mal, la prueba dirige la atención a los temas y las habilidades que se examinan, lo cual tiene implicaciones para los esfuerzos de estudio de los alumnos.

La teoría cognitiva también indica que los estudiantes obtienen beneficios si se les da la oportunidad de demostrar competencia y de trabajar por el aumento de su aprovechamiento, definido mediante criterios compartidos por el maestro, el estudiante y la comunidad (Pellegrino, Baxter, y Glaser, 1999). Como dije antes, brindar a los estudiantes una comprensión clara de los objetivos hace que éstos sean más alcanzables. Y ayudar a los chicos a aprender el significado de los criterios en el contexto de su propio trabajo les ayuda a desarrollar la conciencia metacognitiva sobre qué necesitan para mejorar. La teoría cognitiva no lleva a predecir que dejar de hacer evaluaciones sumativas mejoraría el aprendizaje. De hecho, desde una perspectiva cognitiva, el mejor sistema sería aquel en el que las evaluaciones sumativa y formativa estuvieran mutuamente alineadas con objetivos de aprendizaje orientados conceptualmente, y en el que las evaluaciones sumativas se utilizaran como momentos importantes de logro (quizá reconocidos por la familia y los amigos) después de felices periodos de aprendizaje reforzados por la evaluación formativa.

La evidencia más abrumadora que demuestra los efectos negativos de las prácticas docentes y de calificación proviene de la literatura motivacional. En una revisión comprehensiva de la investigación sobre la motivación en los niños, Wigfield, Eceles, y Rodríguez (1998) relatan el hecho impresionante de que muchos aspectos de la organización en el aula tienen efectos negativos tan generalizados que la creencia de los niños en su competencia, sus metas de realización y logros, su interés en las materias escolares y su motivación intrínseca para aprender, disminuyen conforme los estudiantes van cursando los años de la primaria hasta llegar a segundo de secundaria. Con respecto a las prácticas de calificación, los problemas más importantes tienen que ver con el papel de las calificaciones como recompensas (o castigos), las orientaciones de los estudiantes hacia las metas de desempeño o aprendizaje, y el uso de estándares normativos de evaluación contrapuestos a estándares de evaluación de la maestría o dominio en una materia. Todos estos factores, junto con otros tales como el *locus de control*[9] de los estudiantes y sus

[9] Dimensión utilizada en la teoría de la atribución que se relaciona con la percepción de un competidor de la causa del éxito o fracaso.

sentimientos de competencia interaccionan de maneras complejas. Aquí resumo solamente los patrones más significativos y consistentes.

El uso de calificaciones como recompensa contribuye a lo que Lave y Wenger (1991) llamaron la "comercialización del aprendizaje" (p. 112). Cuando no se da un valor cultural al incremento de la habilidad y la participación de alguien en un esfuerzo, la única razón para participar es obtener un conocimiento superficial que puede exhibirse para que lo evalúen. En revisiones de estudios experimentales, los investigadores descubrieron que el uso de recompensas externas puede minar realmente el interés intrínseco de los estudiantes en una tarea (Deci y Ryan, 1985; Lepper, 1983). Como lo resumió Stipek (1996), las recompensas funcionan para disminuir la motivación intrínseca cuando se perciben como un control y cuando no están directamente relacionadas con un desempeño exitoso. Consistentes con los hallazgos positivos de la investigación sobre la retroalimentación, las recompensas o las alabanzas que transmiten una información positiva sobre la competencia tienen más probabilidades de incrementar la motivación intrínseca.

Los descubrimientos respecto de los efectos de las recompensas están estrechamente relacionados con la investigación sobre el tipo de metas que se proponen alcanzar los estudiantes. Dweck (1986) distingue entre aquellos con *metas de maestría* y aquellos con *metas de desempeño*. Estas disposiciones son independientes de las capacidades académicas de los estudiantes. Aquellos que tienen *metas de maestría* están intrínsecamente motivados, buscan tareas estimulantes y disfrutan las oportunidades de desarrollar nuevas competencias. Es menos probable que tengan miedo a la evaluación porque ven al maestro como un recurso. Cuando los estudiantes con una orientación de maestría se enfrentan a una tarea difícil, es probable que persistan, que mantengan una actitud positiva y que busquen estrategias de solución. Una orientación de maestría también se ha denominado una orientación de tarea por teorías afines. En cambio, los estudiantes con una orientación de metas de desempeño están motivados extrínsecamente. Están más bien interesados en verse competentes que en serlo y tenderán a evitar situaciones en las que pudieran parecer incompetentes. Las teorías relacionadas indican que la orientación del desempeño es provocada por ambientes que fomentan el ego. Cuando encaran una tarea difícil, los estudiantes orientados al desempeño harán con frecuencia comentarios sobre su falta de capacidad, actuarán con aburrimiento o ansiedad, y exhibirán un marcado deterioro en su desempeño. Dweck llamó a estas conductas *impotencia aprendida*. A causa de su miedo a la evaluación, los estudiantes en esta categoría quizá traten de ocultar al maestro su falta de entendimiento.

Es importante hacer notar que la orientación a metas de maestría o de desempeño no son atributos fijos del estudiante; pueden crearse o producirse en diferentes grados por el ambiente de aprendizaje y se han provocado experimentalmente (Elliot y Dweck, 1988). Por ejemplo, en un estudio de estudiantes universitarios, dos grupos desarrollaron niveles muy diferentes de comprensión conceptual según si sabían que les harían una prueba al final del estudio o que necesitarían dar clases a otros sobre el material (Benware y Deci, 1984). Los estudiantes tienen más probabilidades de desarrollar una orientación de aprendizaje cuando los maestros destacan el esfuerzo, el aprendizaje y el trabajo duro, en vez del desempeño y la obtención de la respuesta correcta, cuando los errores se tratan como una parte normal del aprendizaje, y cuando la evaluación del progreso se ve acompañada de oportunidades de mejorar (Ames, 1992; Stipek, 1996).

Tal vez las consecuencias negativas más serias de las prácticas tradicionales de calificación provienen del uso de comparaciones normativas. Como ha indicado Ames (1984), las estructuras de clase competitivas hacen que adquieran

El *locus de control* puede ser interno (es decir, basado en las propias características del competidor, tales como capacidad o esfuerzo) o externos (esto es, debido a factores tales como la suerte, fuera del control del competidor).

importancia las comparaciones sociales y los juicios sobre la capacidad. En una serie de estudios, Butler (1987, 1988), Butler y Nisan (1986) descubrieron que las calificaciones que se distribuían normativamente daban como resultado un interés menor, menos ganas de persistir y un desempeño más bajo en comparación con los estudiantes que habían recibido una retroalimentación sustantiva. En un estudio clásico, Harackiewicz, Abrahams y Wageman (1987) descubrieron que la evaluación que se basaba en normas sociales reducía el interés en una tarea, en tanto que la evaluación basada en lograr un nivel predeterminado incrementaba el interés. La conclusión general de Stipek (1996) respecto de esta literatura fue que la evaluación, especialmente la de tareas difíciles, tiende a minar el interés intrínseco. Sin embargo, la excepción que identificó es digna de atención y anticipa mis recomendaciones en cuanto a las prácticas de calificación:

> No obstante, la evaluación sustantiva que proporciona información sobre competencias y una guía para esfuerzos futuros, y la evaluación que está basada en la maestría o más bien el dominio que en las normas sociales, no tienen, al parecer, estos efectos negativos y pueden incluso aumentar el interés intrínseco en las tareas académicas. (Stipek, 1996, p. 99)

3.5. Parámetros para el desarrollo de la competencia

Para que se respalden mutuamente, las evaluaciones sumativa y formativa deben estar alineadas conceptualmente. Deben ser plenamente capaces de representar objetivos de aprendizaje importantes, y deben usar la misma gama extensa de tareas y de tipos de problemas para representar la comprensión de los estudiantes. Sin embargo, las evaluaciones sumativas no deben ser meras repeticiones de tareas formativas previas, sino que deben ser la culminación de desempeños que inviten a los estudiantes a exhibir su maestría y a utilizar su conocimiento en formas que generalicen y extiendan lo que han aprendido. Las evaluaciones sumativas pueden pensarse como momentos importantes continuos de aprendizaje que apuntalan la evaluación formativa.

Las evaluaciones sumativa y formativa tienen finalidades diferentes. Una hace posible el aprendizaje y, la otra, ilustra la realización y los logros. En vista de los conocidos efectos negativos de la calificación, una pregunta crucial sería la siguiente: ¿cuán a menudo deberíamos recurrir a la evaluación sumativa? Los especialistas en medición defienden la calificación frecuente de trabajos escolares para reunir suficientes datos con objeto de asegurar la confiabilidad. Los cognitivistas consideran que los estudiantes deben tener práctica con los criterios que se usarán para evaluar desempeños culminantes. Sin embargo, el modelo de evaluación formativa y la investigación sobre la motivación sostienen que la calificación podría socavar la orientación de aprendizaje de los estudiantes. Por lo tanto, para conseguir que la evaluación formativa sea realmente para el aprendizaje, tal vez los maestros necesiten posponer otorgar calificaciones, o usarlas sólo cuando el estudiante se auto-evalúe, y como calificaciones *hipotéticas* que ayuden a los estudiantes a permanecer centrados en la retroalimentación sustantiva. Pero lo cierto es que los maestros deben evitar interrumpir y juzgar como si ya estuviera terminada la calidad del aprendizaje que aún está en marcha. Desde luego que la cuestión de la confiabilidad es importante, y no debe calificarse a los estudiantes con base solamente en una o dos pruebas formales aisladas. Sin embargo, si las evaluaciones sumativas están contenidas en las progresiones del aprendizaje, entonces la confiabilidad de los eventos calificados está respaldada por otras evidencias de cada competencia en proceso de desarrollo del estudiante a lo largo de ese continuo fundamental.

Las evaluaciones sumativas y las calificaciones que se basan en ellas deben representar realización y logros. Consistente con el resumen que hice de Stipek (1996), en líneas anteriores, acerca de la investigación sobre motivación; la evaluación del

desempeño debe basarse en estándares de maestría o dominio más que en normas sociales. Las calificaciones basadas en el desempeño se alinearán en forma más transparente a la retroalimentación con los mismos estándares utilizados para la evaluación formativa, y se comunicarán mejor a públicos externos. Cuando los maestros hacen ajustes a las calificaciones para tomar en cuenta el esfuerzo y el mejoramiento, con frecuencia responden a cuestiones de justicia. ¿Es justo juzgar a los estudiantes de capacidades diferentes con los mismos criterios? ¿Y no ocurre que los estudiantes de menores capacidades probablemente dejen de hacer esfuerzos si los estándares están fuera de su alcance? En aulas heterogéneas, calificar desde el punto de vista de los estándares de maestría o dominio requerirá de otros sistemas de respaldo para los estudiantes de diferentes capacidades, incluyendo estrategias tales como: diferente ritmo de aprendizaje y tiempo diferente para las evaluaciones importantes, identificación de metas intermedias y alcanzables, y andamiaje diferencial. Si se toma con seriedad, el compromiso de que las calificaciones representen desempeño significaría suprimir los diversos elementos de la calificación por condescendencia, tales como puntos extra, puntos por entregar fichas y borradores de trabajos, puntos por entregar tareas que jamás se califican y así sucesivamente. Los efectos de las tareas que ayudan a los estudiantes a aprender deben evaluarse en última instancia en evaluaciones culminantes donde el aprendizaje será manifiesto. Al mismo tiempo, estarían permitidas otras formas de ayudar a los estudiantes a suavizar sus preocupaciones sobre las calificaciones, si en cada caso les dieran la oportunidad de que ellos demostraran su maestría. Entre éstas estarían las tareas o pruebas de reemplazo o descartar las calificaciones de exámenes cuando se verifica el aprendizaje mediante evaluaciones posteriores.

4. EVALUACIONES EXTERNAS Y EN GRAN ESCALA

Las evaluaciones nacionales, estatales y distritales se utilizan para reunir datos con el fin de dar respuesta a las preguntas de los hacedores de políticas a cierta distancia del aula. No obstante, en una era de rendición de cuentas basada en pruebas de alto impacto (*high-stakes accountability)*, las pruebas externas pueden tener también profundos efectos en las prácticas del aula. Idealmente, una evaluación externa que estuviera bien alineada con objetivos de aprendizaje ricos desde un punto de vista conceptual, tendría impactos positivos en la enseñanza si ilustrara metas significativas de aprendizaje, pues proporcionaría una retroalimentación útil a los maestros sobre las fortalezas y las debilidades curriculares, y asimismo verificaría logros de estudiantes individuales. Los autores de *Knowing What Students Know* (Pellegrino et al., 2001) imaginaron para el futuro un sistema de evaluación más *equilibrado* y *coherente*, en el cual la evaluación formativa de la clase recibiría igual atención que las pruebas de alto impacto y en la que las evaluaciones en el aula y las externas estarían vinculadas de una manera coherente al mismo modelo fundamental de aprendizaje.

Actualmente, la idealización de *Knowing What Students Know* no se ha realizado. En realidad, un extenso cuerpo de la literatura al respecto ha documentado los efectos negativos en la enseñanza y el aprendizaje (Heubert y Hauser, 1999; Pedulla et al., 2003; Pellegrino et al., 2001; U.S. Congress, *Office of Technology Assessment,* 1992), causados principalmente por los efectos de distorsión que ejerce el enseñar para la prueba, con formatos limitados y una representación muy poco adecuada de los objetivos de aprendizaje significativos.

En un tratamiento más extenso tanto de los efectos positivos como de los negativos de las pruebas de alto impacto (Shepard, Hammerness, Darling-Hammond y Rust, 2005), mis coautores y yo concluimos que existen dos factores que parecen intervenir en la forma en que las pruebas externas remodelan el currículo. El primero es cuán idóneo es el contenido de las pruebas para captar los objetivos del aprendizaje, y el segundo, es la capacidad del docente para mantener en el aula la enseñanza concentrada en el verda-

dero aprendizaje. Con base en estas lecciones de la literatura que investiga cómo se enseña para la prueba, identificamos estrategias que ayudan a mantener enfocada la atención de los estudiantes en el aprendizaje, las cuales son consistentes con la visión de la evaluación formativa de la cultura en el aula. Así, además de la coherencia de contenido entre las pruebas internas y las externas, buscamos preservar una coherencia filosófica entre las actitudes hacia el aprendizaje y la evaluación, las cuales son transmitidas durante las clases normales, y la actitud que uno adopta hacia el aprendizaje cuando se prepara para las pruebas externas.

Para protegerse de los efectos de un currículo guiado por la aplicación de pruebas, los maestros en forma individual, o preferentemente equipos de maestros, deben elaborar una exposición razonada con el fin de ubicar el conocimiento y las habilidades relacionadas con las pruebas dentro de los límites más amplios de los currículos reales. Llamamos a esta técnica *mapeo del campo*. Si comienzan con el marco de referencia del currículo estatal o los estándares nacionales de contenido, los maestros pueden dibujar un diagrama de Venn o construir una tabla para ilustrar cuál subparte del currículo deseado es abarcado por la prueba y cuál no. Muchas pruebas elaboradas comercialmente abarcan la parte más fácil de medir de cada elemento del contenido de un campo, pero esto no significa que se represente el campo adecuadamente. Señalar lo que se ha dejado fuera ayuda a esclarecer las limitaciones de la prueba como guía curricular. Con base en este análisis explícito, los maestros pueden entonces planear conscientemente unidades de estudio y distribución del tiempo de enseñanza en formas que mantengan la atención al contenido de las pruebas en su lugar proporcional. Por consiguiente, no se permitiría que las pruebas de habilidades básicas y de bajo nivel, que representan sólo una subparte de los estándares de contenido deseados, tuvieran tanta influencia en la enseñanza dentro del aula como las evaluaciones basadas en conceptos, que son más estimulantes.

Si estamos de acuerdo con la idea de que los estudiantes deben estar conscientes de su propio progreso en el aprendizaje, también tiene sentido ayudarles a entender la relación entre el conocimiento que se requiere para las pruebas y cómo se compara esto con las formas en que usan su conocimiento en otros ambientes. Por ejemplo, Calkins, Montgomery y Santman (1998) proponen que enseñemos a los niños cómo *leer* las pruebas, igual que les enseñamos a dominar las características de cualquier otro género. "Si nuestros pequeños están acostumbrados a reunirse en la alfombra a leer juntos un texto, que quizá es más largo de lo normal, y luego a hablar de las estrategias para abordarlo o para manejar las dificultades que plantea, ¿por qué no pueden hacer lo mismo con el extracto de una prueba estandarizada de lectura?" (p. 71). De igual manera, en el estudio de McNeil (1988), maestros experimentados de escuelas *magnet*[10] buscaron la forma de contrarrestar el enfoque de *fragmentos y hechos* que tenían las pruebas y ayudaron a sus alumnos a llevar dos conjuntos de notas, uno para el conocimiento real y otro para el conocimiento que necesitarían para la prueba. Si bien los estudiantes, especialmente los de primaria, merecen tener algo de práctica con formatos de pruebas que les piden hacer ejercicios de respuesta abierta (Matemáticas), o redactar párrafos a partir de una instrucción general, para que no se desconcierten ante tareas poco familiares, tal práctica debe darse en el contexto de los objetivos de la enseñanza con los que estas tareas se relacionan. Y en vista de lo que sabemos sobre la falta de transferencia cuando sólo se utiliza un tipo de problema, los maestros pueden trabajar para asegurarse de que haya una comprensión más sólida si se concentran en los principios fundamentales y si continúan pidiendo a sus alumnos que hagan extensiones y aplicaciones de lo aprendido.

[10] Las *magnet schools* son escuelas públicas que atraen a estudiantes de otro vecindario para reducir o eliminar el desequilibrio racial. Estas escuelas dan una importancia especial al logro académico o a un campo en particular, como Ciencias, Matemáticas, Artes o Ciencias de la computación. [N. T.]

Buena parte de este consejo, sobre cómo mantener nuestra integridad profesional ante las presiones de las pruebas de alto impacto, tiene que ver con la actitud y la posición que uno adopte. La idea fundamental es poner atención en la prueba sólo en la medida en que ésta se relacione con el currículo, más que detenerse en la prueba y dejar que ésta se convierta en el centro de la planeación de la enseñanza. Tal posición permite a los maestros utilizar los resultados de las pruebas para hacer sólo las mejoras apropiadas y necesarias en el currículo y la enseñanza. Como lo señaló Wiggins (1998), los buenos maestros tienen la capacidad de auto-evaluarse, pero aún así hay puntos ciegos y una carencia de referentes externos. La mayoría de los maestros con un buen conocimiento de los estándares de contenido (puesto que la prueba se refiere a ellos), y con una idea bastante clara de lo que sus alumnos saben, pueden predecir cómo saldrán éstos. Por lo tanto, las aportaciones más importantes de los resultados de las pruebas externas pueden a menudo venir de los que resultan sorpresivos, ya que los maestros se ven llevados a preguntarse cosas como éstas: ¿por qué no fueron capaces mis alumnos de resolver este problema (o de realizar esta tarea)? ¿Cómo debo cambiar mi enseñanza para ayudarles a resolver problemas de este tipo? ¿Por qué no supe lo que ellos no sabían? Además, aun cuando los resultados concuerden con sus predicciones, los maestros pueden evaluar la idoneidad de sus propios esfuerzos si examinan las relativas fortalezas y debilidades de los elementos curriculares. Por ejemplo, ¿se desempeñan los estudiantes igualmente bien en las habilidades de pensamiento de orden más elevado que en las habilidades básicas? Cuando los maestros tienen un conocimiento profesional bien desarrollado sobre cómo encaja la información basada en pruebas en un marco de referencia más amplio de objetivos curriculares enriquecidos, se hace posible prestar atención a una comprensión proveniente de los resultados de exámenes, sin temor a reducir y distorsionar el currículo, tan bien documentados en la literatura.

5. CONCLUSIONES: IMPLICACIONES PARA LA INVESTIGACIÓN Y LA TEORÍA DE LA MEDICIÓN

Ralph Tyler es un icono en la historia de la medición educativa. Empecé este capítulo con un análisis histórico porque yo quería abordar explícitamente la tensión entre mi constante creencia en la visión de Tyler —de que la medición educativa debe ser parte integral de la enseñanza— y mi necesidad de cuestionar su herencia involuntaria: el persistente modelo de pruebas y mediciones para la evaluación en el aula, fundado principalmente en pruebas objetivas elaboradas externamente. La otra visión de la evaluación en el aula presentada en este capítulo es radicalmente diferente de la imagen de la aplicación de pruebas en el salón de clase presentada en volúmenes anteriores de *Educational Measurement,* y en libros de texto sobre medición. Ésta otra visión se centra en conceptualizaciones mucho más ricas sobre el aprendizaje del estudiante en el contexto de actividades significativas y resalta el uso formativo de la evaluación para mejorar el aprendizaje.

El nuevo modelo idealizado está bien fundado en las teorías cognitivas y socioculturales contemporáneas del aprendizaje y en la teoría de la motivación, pero en cierto sentido todavía es experimental, ya que no se ha puesto en marcha en gran escala. Queda mucho por aprender sobre qué nuevas ideas y adaptaciones adicionales se necesitarán para hacer que tal visión funcione en la práctica. Concluyo reflexionando sobre cuatro temas esenciales para la investigación futura: (1) estudios de las herramientas y los procesos de evaluación, (2) desarrollo de las progresiones del aprendizaje, (3) estudios del desarrollo del maestro, y (4) nuevas conceptualizaciones de confiabilidad y la validez.

5.1. Estudios de las herramientas y los procesos de evaluación

En *Knowing What Students Know*, Pellegrino *et al.* (2001) sostuvieron que las interpretaciones

y las prácticas de evaluación deben basarse en un modelo bien concebido de aprendizaje del estudiante. Lo mismo puede decirse respecto de la importancia del modelo teórico que sirve de fundamento a los estudios de investigación. Cuando Black y Wiliam (1998) emprendieron su voluminosa revisión de la investigación sobre la evaluación formativa en el aula, citaron estudios fundamentales, meta-análisis y revisiones que representaban mucho más de mil estudios. Con todo, muchos de los estudios citados son inadecuados para nuestros propósitos, porque se derivan de modelos de aprendizaje muy insuficientes para la teoría contemporánea. Por ejemplo, una gran mayoría de los estudios sobre retroalimentación se conceptualizaron desde una perspectiva conductista y dependían de pre-pruebas y post-pruebas que se parecían mucho a los materiales de enseñanza. En la literatura motivacional, la recomendación de evitar tareas estimulantes tal vez no se sostenga en un ambiente basado en criterios más que en uno que remita a normas, o cuando la cultura del aula respalda a los estudiantes para que desarrollen una orientación hacia el aprendizaje. Se necesitan nuevos estudios que reflejen modelos constructivistas y socioculturales del aprendizaje asistido.

Unos cuantos estudios citados en el capítulo ilustran cómo puede introducirse y estudiarse una práctica específica de evaluación, consistente con la teoría social-constructivista. Por ejemplo, Elawar y Corno (1985) crearon una intervención para ayudar a los maestros a que se concentraran en conseguir que la retroalimentación fuera útil para el mejoramiento de los estudiantes. White y Frederiksen (2000) y Klenowski (1995) examinaron el efecto de la auto-evaluación en el aprendizaje del estudiante y sus actitudes respecto a las calificaciones, el conocimiento de los criterios, etc. Se necesitan más estudios como éstos. Deben diseñarse para que respondan a preguntas de investigación tales como las siguientes: ¿cuán bien captan objetivos importantes de aprendizaje las evaluaciones incorporadas en la enseñanza? ¿Qué conocimientos aportan tales evaluaciones a los maestros y estudiantes, y qué otros estímulos proporcionan para que se hagan avances en el aprendizaje? ¿Cómo puede hacerse que las ocasiones de la evaluación y las estrategias para realizarla se adapten a las rutinas de enseñanza? (Ejemplos de estas estrategias son sobre todo los portafolios, las discusiones entre estudiantes y la explicación a la clase de las soluciones a problemas.) ¿Pueden introducirse con muchísimo cuidado los procesos derivados de la teoría del aprendizaje —evaluación del conocimiento previo, retroalimentación, conciencia metacognitiva, auto-evaluación, etcétera— como un medio para incrementar el aprendizaje y la motivación del estudiante? ¿En qué condiciones y para quién son eficaces estas estrategias? ¿Pueden hacerse coherentes, y que se respalden mutuamente la evaluación formativa y la sumativa? Cuando son coherentes, ¿se incrementan el aprendizaje y la motivación del estudiante? Las extrapolaciones especulativas de la investigación y la teoría existentes también deben someterse a prueba explícitamente. Por ejemplo, si se pospusieran las calificaciones, ¿aumentaría esto la eficacia de la evaluación formativa, es decir, se vería estimulada la orientación al aprendizaje y aumentarían la realización y los logros?

5.2. Estudios del desarrollo del maestro

Los estudios de una variable por vez o de una característica de la evaluación por vez son útiles para los fines de la investigación, porque nos ayudan a concentrarnos en los efectos de una estrategia específica y en las formas mediante las cuales esas prácticas de evaluación en particular respaldan el aprendizaje del estudiante. Además, estos estudios me parecen ejemplos útiles de cómo puede ocurrir el cambio: es decir, cómo es posible que los maestros a quienes les gusta la teoría de la evaluación formativa den los pasos iniciales para modificar su práctica. En última instancia la meta es hacer un paradigma completo y un cambio cultural (Shepard, 2000). Sin duda, si uno adopta una *técnica* de evaluación formativa sin un cambio filosófico correspondiente,

los esfuerzos se verán socavados o acabarán por ser inútiles, ya que persistirán las actitudes tradicionales, como la de quienes en su desempeño académico sólo les importa la calificación. Sin embargo, sabemos gracias a la investigación sobre la reforma educativa y el cambio del docente que es imposible instaurar una reforma integral. Por lo tanto, es bastante práctico tomar en consideración argumentos teóricos y luego adoptar una estrategia específica de evaluación, como el punto focal para la transformación lenta y concienzuda en la práctica, junto con un acompañamiento constante y la reflexión para tener en cuenta las inevitables repercusiones teóricas y prácticas. La estrategia de la intervención única de evaluación podría ser una capacitación especial para dar retroalimentación enfocada al mejoramiento; documentar los fondos del mismo conocimiento de los estudiantes; examinar el trabajo de los alumnos en equipos de grado; introducir la auto-evaluación, y así sucesivamente. En un estudio notable realizado en Inglaterra en el que el avance promedio de los alumnos de un grupo fue del orden de un tercio de una desviación estándar, Black y Wiliam (2004) familiarizaron a los maestros con la teoría de la evaluación formativa y luego los invitaron a desarrollar sus propios planes, seleccionando estrategias de la literatura, tales como preguntas abundantes, marcar el trabajo escolar sólo con comentarios, compartir criterios con los alumnos e implantar la auto-evaluación y la evaluación por pares dada por los estudiantes.

La investigación sobre el aprendizaje y desarrollo profesional de los maestros nos proporciona varios principios generales para respaldar el cambio. En esencia, nosotros (los que promovemos el cambio) necesitamos tratar a los maestros en la medida en que son personas que aprenden, de la misma manera como les pedimos que traten a sus estudiantes. Necesitamos tener un modelo bien concebido de la práctica profesional ideal hacia la que los conducimos, y al mismo tiempo debemos estar conscientes de que los maestros harán contribuciones y también modificarán las herramientas y prácticas de la comunidad en la que participan (Lave y Wenger, 1991). Para que los docentes hagan cambios significativos en las creencias pedagógicas y en las prácticas concomitantes, ellos mismos necesitarán experimentar y reflexionar sobre nuevos procedimientos en el contexto de sus propias aulas (Putnam y Borko, 2000). Los estudios de investigación realizados en el contexto de la práctica deben prestar atención a cuestiones tales como las siguientes: ¿qué condiciones, creencias previas o apoyos hacen posible o frustran el uso de la evaluación tal como uno la quiere hacer? ¿Cómo influye el conocimiento de la materia que tiene un maestro sobre sus creencias y sobre la implementación de prácticas eficaces de evaluación formativa? ¿Cómo pueden quedar integradas las prácticas de evaluación formativa con otras reformas curriculares o con otros cambios culturales dirigidos a desarrollar una comunidad de personas que aprenden, los cuales también se basan en un modelo sociocultural del aprendizaje? ¿Cómo debe uno dedicarse a las cuestiones en torno a la calificación y a las evaluaciones externas, con objeto de ayudar y no de obstaculizar los esfuerzos de la evaluación formativa? ¿Cómo influyen en la renegociación de los objetivos de la evaluación formativa el contexto escolar y el contrato social implícito (Perrenoud, 1991) que los estudiantes llevan consigo en lo que se refiere a las calificaciones?

5.3. Nuevas conceptualizaciones de la confiabilidad y la validez

Mi preámbulo histórico describió el campo de la medición como reacio a distanciarse de una visión tradicional de las pruebas en el aula, que se concentraban en la evaluación sumativa. La llegada del siglo XXI, sin embargo, ha visto cambios notables. La publicación de *Knowing What Students Know* (Pellegrino et al., 2001) fue un verdadero parteaguas. Un segundo acontecimiento muy relevante fue la publicación de un número especial de *Educational Measurement: Issues and Practice* sobre "cambiando la manera en que los teóricos de la medición piensan sobre

las evaluaciones en el aula" (Brookhart, 2003). Al frente de este movimiento, Brookhart (2003) sostiene que los avances en la teoría de la confiabilidad y la validez han sido forjados en el contexto de programas de evaluación en gran escala. En atención a la necesidad de que evaluación y enseñanza se integren en el aula, como he señalado en este capítulo, Brookhart (2003) ha dado argumentos en el sentido de que la forma de concebir confiabilidad y validez en el aula debe ser también fundamentalmente diferente.

Un principio básico de la teoría de la medición ha sido siempre que la confiabilidad y la validez dependen del uso de pruebas. En el contexto del aula, no es necesario que la confiabilidad de cualquier evaluación cumpla el mismo criterio de estabilidad que una medición utilizada para determinar la graduación del bachillerato o el ingreso a la universidad. Debido a que la evaluación formativa en las aulas es constante, una percepción errónea de las habilidades o el conocimiento de un estudiante puede un día corregirse con nueva información y al siguiente con una demostración de su aprovechamiento (Shepard, 2000). Lo que es más importante, los conceptos de andamiaje y evaluación dinámica (Lidz, 1987), fundamentados en la zona de desarrollo próximo de Vygotsky, tienen la finalidad de cambiar el nivel de aprovechamiento del estudiante en el mismo proceso de evaluación. Moss (2003), en el número especial de *Educational Measurement: Issues and Practice*, sostiene que, como docente, "No tengo necesidad de sacar y garantizar interpretaciones fijas de las capacidades de los estudiantes, sino que más bien mi trabajo es ayudarles a hacer que esas interpretaciones se vuelvan obsoletas" (p. 16). Smith (2003), también en el número especial, propuso que quizás la *suficiencia de la información* sería el criterio más apropiado para la confiabilidad en el contexto del aula. "¿Tengo a la mano la información suficiente para tomar una decisión razonable respecto de este estudiante en relación con este campo de contenido?" (p. 30). Como observó Moss, el criterio sería diferente dependiendo de si la decisión fuera sumativa, como para una carta de recomendación, o formativa, como cuando un maestro adapta la retroalimentación para ayudar a un estudiante a mejorar el trabajo que ha escrito.

Así como la evaluación formativa quizás no necesite que se dé a conocer una *calificación* fija, también es el caso que los coeficientes de correlación rara vez pueden ser los indicadores apropiados de la confiabilidad. En las aulas, dar sentido a los datos de observación y a las muestras del trabajo de los alumnos significa buscar patrones, comprobar evidencia contradictoria y comparar la descripción emergente en contraposición con modelos del desarrollo de las competencias. Al igual que otros autores, estoy en favor de un enfoque interpretativista del análisis y la síntesis de datos (Gipps, 1999; Graue, 1993; Moss, 1996, 2003; Shepard, 2001). En mi propio caso, veo una fuerte conexión entre el uso de métodos cuantitativos de investigación, de prácticas de evaluación formativa, y mi preparación para el trabajo clínico, cuando utilizo observaciones para formar una hipótesis tentativa, cuando reúno información adicional para confirmar o revisar, y cuando planeo una intervención (que en sí misma es una hipótesis de trabajo). En efecto, hace un tiempo, Geertz (1973) estableció una analogía entre la inferencia clínica como la que se usa en medicina y la forma en que los teóricos culturales *diagnostican* el significado fundamental del discurso social, queriendo decir con ello que usan una *teoría aterrizada* (*grounded*) para generar interpretaciones convincentes, o generalizaciones que tienen un poder explicativo más allá de las descripciones gruesas. Obsérvese que verificar si existen patrones y justificar que ciertas interpretaciones particulares están fundamentadas borra los límites entre la confiabilidad y la validez, como Smith (2003) tuvo que reconocer en su argumento en favor de la suficiencia de la información como la forma de definir la confiabilidad concerniente a los propósitos del aula. De hecho, la línea de interrogación que Smith (2003) propone para evaluar la confiabilidad corresponde de cerca al proceso interpretativo de evidencia que Moss

(2003) ofrece como criterio de validez. En vista de que esta última es el concepto más incluyente, prefiero usar el término validez para referirme al proceso de fundamentar interpretaciones, y limitar el uso del término confiabilidad a las exigencias de consistencia más estrechas, tales como los acuerdos entre varios jueces al utilizar guías de calificación (*rubrics*).

Brookhart (2003), Moss (2003) y Smith (2003) han comenzado a especular sobre cómo tendrá que reformularse la teoría de la confiabilidad y la validez para que tengan sentido en los propósitos del aula, pero es tan sólo el inicio de un esfuerzo mayor de reconceptualización. Será necesario modificar muchos viejos conceptos o incluso ir en contra de ellos. Por ejemplo, sabemos que la diferencia entre dos calificaciones cuantitativas, es por lo general mucho menos confiable que cualquiera de las dos calculada por sí sola. No obstante, si fuéramos a tomar con seriedad la recomendación de *Knowing What Students Know*, de representar el progreso del estudiante a lo largo de continuos de aprendizaje bien desarrollados, las *calificaciones de diferencia* o las mediciones del progreso podrían ser sumamente confiables, porque la medición de crecimiento basada en el aula tendría el respaldo de todos los datos que se reunieran a lo largo del proceso, y no dependería solamente de los dos extremos. Un programa de investigación que estudiara la confiabilidad para los propósitos del aula debería empezar por examinar el tipo y el grado de consistencia requerido para respaldar las decisiones en el aula.

La validez es el sitio apropiado para empezar o terminar cualquier intento de medición educativa. En las aulas, la evaluación formativa es válida si contribuye al progreso del aprendizaje del estudiante. Como señala Moss (2003), la validez en los contextos del aula alude principalmente a las consecuencias, a qué tan bien las interpretaciones de las evaluaciones informan a las decisiones docentes y cuánto ayudan a hacer que los estudiantes avancen a lo largo de una trayectoria de competencia creciente. Para trazar un plan de investigación con objeto de desarrollar una teoría de la validez apropiada para la evaluación en el aula, uno necesita solamente retomar los temas importantes del capítulo. Para que se respalden mutuamente, la evaluación formativa y la sumativa deben alinearse desde el punto de vista conceptual. Deben incorporar objetivos de aprendizaje importantes, que se sirvan de una amplia gama de tareas y tipos de problemas para captar las comprensiones de los estudiantes. Las investigaciones sobre la validez deben examinar lo bien que diversas herramientas de evaluación representan el conocimiento, las habilidades y las actitudes de los estudiantes así como sus identidades en proceso de desarrollo, con fines sumativos, y cuán bien soportan acercamientos profundos y de dominio del aprendizaje, cuando se usan formativamente. Los estudios de validez deben, asimismo, consagrarse a revisar si los procesos de evaluación funcionan como se desea. Por ejemplo, ¿utilizan los estudiantes una retroalimentación sustantiva para mejorar su trabajo? El programa de investigación esbozado líneas arriba para examinar los efectos de las herramientas y los procesos de evaluación es, en realidad, un programa de investigación sobre la validez. La mayor comprensión que se alcance gracias a la investigación que pone a prueba el modelo de evaluación formativa, su relación con la teoría del aprendizaje, etcétera, ayudará asimismo a iluminar una nueva comprensión de la teoría de la validez, necesaria para la evaluación en el aula.

Bibliografía

Ames, C. (1984). Competitive, cooperative, and individualistic goal structures: A cognitive-motivational analysis. En R. E. Ames y C. Ames (Eds.). *Research on motivation in education*, 1, 177-207. Nueva York: Academic Press.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84, 261-271.

Assessment Reform Group. (1999). *Assessment for learning: Beyond the black box.* University of Cambridge School of Education: Cambridge.

Atkin, J. M., Black, P., y Coffey, J. (2001). *Classroom assessment and the National Science Education Standards.* Washington, DC: National Academy Press.

Baron, J. B., y Wolf, D. P. (Eds.). (1996). *Performance-based student assessment: Challenges and Possibilities, Ninety-fifth Yearbook of the National Society for the Study of Education.* Parte 1. Chicago: University of Chicago Press.

Benware, C., y Deci, E. (1984). Quality of learning with an active versus passive motivational set. *American Educational Research Journal*, 21, 755-765.

Black, P., y William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7-74.

Black, P., y William, D. (2004). The formative purpose: Assessment must first promote learning. En M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103° Yearbook of the National Society for the Study of Education.* Parte 2, pp. 20-50. Chicago: University of Chicago Press.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals.* Nueva York: David McKay Company.

Bransford, J. D., Brown, A. L., y Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school.* Washington, DC: National Academies Press.

Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30,* 123-142.

Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education, 7* (4), 279-301.

Brookhart, S. M. (2003) Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice, 22* (4), 5-12.

Brown, F. G. (1981). *Measuring classroom achievement.* Nueva York: Holt, Rinchart y Winston.

Bruner, J.S. (1985). Vygotsky: A historical and conceptual perspective. En J. V. Wersch y Center for Psychosocial Studies (Eds.), *Culture, communication, and cognition: Vygotskian perspectives,* pp. 21-34. Nueva York: Cambridge University.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest and performance. *Journal of Educational Psychology, 79,* 474-482.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58,* 1-14.

Butler, R., y Nisan, M. (1986). Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance. *Journal of Educational Psychology, 78,* 210-216.

Calkins, L., Montgomery, K., y Santman, D. (1998). *A teacher's guide to standardized reading tests: Knowledge is power,* Portsmouth, NH: Heinemann.

Carpenter, T. P., y Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education. 15* (3), 179-202.

Cizek, G. J., Fitzgerald, S. M., y Rachor, R. E. (1995/1996). Teachers' assessment prac-

tices: Preparation, isolation, and the kitchen sink. *Educational Assessment, 3* (2), 159-179.

Clay, M. M. (1985). *The early detection of reading difficulties.* Auckland, Nueva Zelanda: Heinemann.

Cobb, P., Wood, T., y Yackel, E. (1993). Discourse, mathematical thinking, and classroom practice. En E. A. Forman, N. Minick, y C. A. Stone (Eds.), *Contexts for learning: Sociocultural dynamics in children's development,* pp. 91-119. Nueva York: Oxford University Press.

Connolly, A., Nachtman, W., y Pritchett, M. (1972). *Keymath diagnostic arithmetic test.* Circle Pines, MN: American Guidance Service.

Cook, W. W. (1941). Achievement tests. En W. S. Monroe (Ed.), *Encyclopedia of educational research,* pp. 1283-1301. Nueva York: Macmillan.

Cook, W. W. (1951). The functions of measurement in the facilitation of learning. In E. F. Lindquist (Ed.), *Educational Measurement*, pp. 3-46. Washington, DC: American Council of Education.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58* (4), 428-481.

Deci, E., y Ryan, R. (1985). *Intrinsic motivation and self-determination in human behavior.* Nueva York: Plenum.

Dorr-Bremme, D. W. (1983). Assessing students: Teachers' routine practices and reasoning. *Evaluation Comment, 6* (4), 1-12.

Dweek, C. (1986). Motivational processes affecting learning. *American Psychologist, 41,* 1040-1048.

Elawar, M. C., y Corno, L. (1985). A factorial experiment in teachers' written feedback on student homework: Changing teacher behavior a little rather than a lot. *Journal of Educational Psychology, 77,* 162-173.

Elliot, E., y Dweck, C. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54,* 5-12.

Farr, R., y Griffin, M. (1973). Measurement gaps in teacher education. *Journal of Research and Development in Education, 7* (1), 19-28.

Fennema, E., y Franke, M. L. (1992). Teachers' knowledge and its impact. En D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning,* pp. 147-164. Nueva York: Macmillan.

Fleming, M., y Chambers, B. (1983). Teacher-made tests: Windows on the classroom. *New Directions for Testing and Measurement: Testing in the Schools, 19,* 29-38.

Forster, M., y Masters, G. (2004). Bridging the conceptual gap between classroom assessment and system accountability. En M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103° Yearbook of the National Society for the Study of Education,* parte 2, pp. 51-73. Chicago: University of Chicago Press.

Geertz, C. (1973). *The interpretation of cultures.* Nueva York: Basic Books.

Gipps, C. V. (1999). Socio-cultural aspects of assessment. En P.D. Pearson y A. Iran-Nejad (Eds.), *Review of research in education, 24*, 355-392. Washington, DC: American Educational Research Association.

Glaser, R., y Nitko, A. J. (1971). Measurement in learning and instruction. En R. L. Thorndike (Ed.), *Educational measurement,* 2ª ed., pp. 625-670. Washington, DC: American Council on Education.

Goehring, H. J., Jr. (1973). Course competencies for undergraduate courses in educational tests and measurement. *The Teacher Educator, 9,* 11-20.

Goodman, Y. M. (1985). Kidwatching: Observing children in the classroom. En A. Jaggar y M. T. Smith-Burke (Eds.), *Observing the language learner,* pp. 9-18. Newark, DE: International Reading Association and National Council of Teachers of English.

Goslin, D. A. (1967). *Teachers and testing.* Nueva York: Russell Sage.

Graue, M. E. (1993). Integrating theory and practice through instructional assessment. *Educational Assessment, 1,* 293-309.

Harackiewicz, J., Abrahams, S., y Wageman, R. (1987). Performance evaluation and intrinsic motivation: The effects of evaluative focus, rewards, and achievement orientation. *Journal of Personality and Social Psychology, 53,* 1015-1023.

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms.* Cambridge: Cambridge University Press.

Heubert, J., y Hauser, R. (Eds.). (1999). *High-stakes: Testing for tracking, promotion, and graduation.* Washington, DC: National Academy Press.

Hiebert, E. H., y Raphael, T. E. (1998). *Early literacy instruction.* Fort Worth, TX: Harcourt Brace College Publishers.

Hogan, K., y Pressley, M. (1997). Scaffolding scientific competencies within classroom communities of inquiry. En K. Hogan y M. Pressley (Eds.), *Scaffolding student learning: Instructional approaches and issues*, pp. 74-107. Cambridge, MA: Brookline Books.

Klenowski, V. (1995). Student self-evaluation process in student-centered teaching and learning contexts of Australia and England. *Assessment in education, 2,* 145-163.

Kluger, A. N., y DeNisi, A. (1996). The effect of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254-284.

Kulm, G. (Ed.). (1990). *Assessing higher order thinking in mathematics.* Washington, DC: American Association for the Advancement of Science.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S. N., y Phillips, E. D. (1998). *Connected mathematics, say it with symbols: Algebraic reasoning.* Student Ed. Menlo Park, CA: Dale Seymour Publications.

Lave, J., y Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge: Cambridge University Press.

Lepper, M. (1983). Extrinsic reward and intrinsic motivation: Implications for the classroom. En J. Levine y M. Wang (Eds.), *Teacher and student perceptions: Implications for learning,* pp. 281-317. Hillsdale, NJ: Lawrence Erlbaum.

Lidz, C. S. (1987). *Dynamic assessment: An interactional approach to evaluating learning potential.* Nueva York: Guilford Press.

Lindquist, E. F. (Ed.). (1951). *Educational measurement.* Washington, DC: American Council on Education.

Linn, R. L. (Ed.). (1989). *Educational measurement,* 3ª ed. Nueva York: American Council on Education y Macmillan.

Lorsbach, A. W., Tobin, K., Briscoe, C., y LaMaster, S. U. (1992). An interpretation of assessment methods in middle school science. *International Journal of Science Education, 14* (3), 305-317.

Madaus, G. F., y Stufflebeam, D. L. (2000). Program evaluation: A historical overview. En D. L. Stufflebeam, G. F. Madaus, y T. Kellaghan (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation,* 2ª ed., pp. 3-18. Boston: Kluwer Academic.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., y Viator, K. A. (1992, October). *The influence of testing on teaching math and science in grades 4-12: Executive summary.* Chestnut Hill, MA: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

Marshall, H. H. (1988). Work or learning: Implications of classroom metaphors. *Educational Researcher, 17* (9), 9-16.

Marton, F. y Saljo, R. (1976). On qualitative differences in learning: 1. Outcome and process. *British Journal of Educational Psychology, 46,* 4-11.

Masters, G., y Forster, M. (1997). *Mapping literacy achievement: Results of the 1996 National School English Literacy Survey.* Camberra, Australia: Department of Employment, Education, Training and Youth Affairs (DEETYA).

Mathematical Sciences Education Board. (1993). *Measuring what counts: A conceptual guide for mathematics assessment.* Washington, DC: National Academy Press.

Mayo, S. T. (1964, February). *What experts think teachers ought to know about educational measurement.* Ponencia presentada en la reunion anual del National Council on Measurement in Education, Chicago.

Mazur, E. (1997). *Peer instruction: A user's manual.* Upper Saddle River, NJ: Prentice Hall.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice, 22* (4), 34-43.

McNeil, L. M. (1988). *Contradictions of control: School structure and school knowledge.* Nueva York: Routledge.

Mitchell, M. (1993). Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology, 85,* 424-436.

Moll, L.C., Amanti, C., Neff, D., y González, N. (1992). Funds of knowledge for teaching: Using a qualitative approach to connect homes and classrooms. *Theory Into Practice, 31* (2), 132-141.

Morrow, L. M. (1985). Retelling stories: Strategies for improving children's comprehension, concept of story structure, and oral language complexity. *Elementary School Journal, 85,* 647-661.

Morrow, L. M., y Smith, J. K. (1990). *Assessment for instruction in early literacy.* Englewood Cliffs, NJ: Prentice Hall.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher, 25,* 20-28, 43.

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice, 22* (4), 13-25.

National Association for the Education of Young Children. (1990). *Guidelines for appropriate curriculum content and assessment in programs serving children ages 3 through 8.* Washington, DC: National Association for the Education of Young Children.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Autora.

National Council of Teachers of Mathematics. (1995). *Assessment Standards for School Mathematics.* Reston, VA: Autora.

National Research Council. (1989). *Everybody counts: A report to the nation on the future of mathematics education.* Washington, DC: National Academy Press.

Nava, F. J. G., y Loyd, B. H. (abril de 1992). *An investigation of achievement and non-achievement criteria in elementary and secondary school grading.* Ponencia presentada en la reunion anual de la *American Educational Research Association*: San Francisco.

Newmann, F. (1992). Higher order thinking and prospects for classroom thoughtfulness. En F. Newmann (Ed.), *Student engagement and achievement in American secondary schools,* pp. 62-91. Nueva York: Teachers College Press.

Nitko, A. J. (1989). Designing tests that are integrated with instruction. En R. L. Linn (Ed.), *Educational measurement,* 3ra ed., pp. 447-474. Nueva York: American Council on Education y Macmillan.

Noll, V. H. (1955). Requirements in educational measurement for prospective teachers. *School and Society, 80,* 88-90.

Ogle, D. M. (1986). K-W-L: A teaching model that develops active reading of expository text. *The Reading Teacher, 39* (6), 564-570.

Pedulla, J. J., Abrams, L. M., Madaus, G. F., Russell, M. K., Ramos, M. A., y Miao, J. (2003). *Perceived effects of state-mandated testing programs on teaching and learning: Findings from a national survey of teachers.* Boston: Boston College, National Board on Educational Testing and Public Policy.

Pellegrino, J. W., Baxter, G. P., y Glaser, R. (1999). Addressing the "Two Disciplines" problem: Linking theories of cognition

and learning with assessment and instructional practice. En P. D. Pearson y A. Iran-Nejad (Eds.), *Review of research in education,* 24, 307-353. Washington, DC: American Educational Research Association.

Pellegrino, J. W., Chudowsky, N., y Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Perrenoud, P. (1991). Towards a pragmatic approach to formative evaluation. En P. Weston (Ed.), *Assessment of pupils' achievement: Motivation and school success,* pp. 77-101. Amsterdam: Swets and Zeitlinger.

Putnam, R. T., y Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher, 29* (1), 4-15.

Resnick, L. B., y Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for education reform. En B. R. Gifford y M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction,* pp. 37-75. Boston: Kluwer Academic.

Roeder, H. H. (1972). Are today's teachers prepared to use tests? *Peabody Journal of Education, 59,* 239-40.

Romberg, T., y Carpenter, T. (1986). Research on teaching and learning mathematics: Two disciplines of scientific inquiry. En M. Wittrock (Ed.), *Handbook of research on learning,* 3ª ed., pp. 850-873. Nueva York: Macmillan.

Routman, R. (2000). *Conversations: Strategies for teaching, learning and evaluating.* Portsmouth, NH: Heinemann.

Sadler, R. (1989). Formative assessment and the design of instructional assessments. *Instructional Science, 18,* 119-144.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening? *Phi Delta Kappan, 69*(9), 631-634.

Seriven, M. (1967). The methodology of evaluation. En R. A. Tyler, R. M. Gagne, y M. Seriven (Eds.), *Perspectives of curriculum evaluation,* pp. 39-83. Chicago: Rand McNally.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher; 29* (7), 4-14.

Shepard, L. A. (2001) The role of classroom assessment in teaching and learning. En V. Richardson (Ed.), *Handbook of research on teaching,* 4ª ed., pp. 1066-1101. Washington, DC: American Educational Research Association.

Shepard, L. A. (2003). Reconsidering large-scale assessment to heighten its relevance to learning. En J. M. Atkin y J. E. Coffey (Eds.) *Everyday assessment in the science classroom,* pp. 121-146. Arlington, VA: NSTA Press.

Shepard, L. A., y Bliem, C. L. (1995). Parents' thinking about standardized tests and performance assessments. *Educational Researcher, 24* (8), 25-32.

Shepard, L., Hammerness, K., Darling-Hammond, L., y Rust, F. (2005). Assessment. En L. Darling-Hammond y J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do,* pp. 275-326. San Francisco: Jossey-Bass.

Silver, E. A., y Kenney, P. A. (1995). Sources of assessment information for instructional guidance in mathematics. En T. A. Romberg (Ed.), *Reform in school mathematics and authentic assessment,* pp. 38-86. Albany, NY: State University of New York Press.

Silver, E. A., y Kilpatrick, J. (1989). Testing mathematical problem solving. En R. Charles y E. Silver (Eds.), *Teaching and assessing mathematical problem solving,* pp. 178-186. Hillsdale, NJ: Lawrence Erlbaum.

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational measurement: Issues and Practice, 22* (4), 26-33.

Stiggins, R. J. (1991). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice, 10* (1), 7-12.

Stiggins, R. J. (2001). *Student-involved classroom assessment,* 3ª ed. Upper Saddle River, NJ: Prentice Hall.

Stiggins, R. J., y Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment.* Albany: State University of New York Press.

Stipek, D. J. (1996). Motivation and instruction. En D. C. Berliner y R. C. Calfee (Eds.), *Handbook of educational psychology,* pp. 85-113. Nueva York: Simon & Schuster Macmillan.

Stipek, D. (1998). *Motivation to learn: From theory to practice,* 3ª ed.,. Boston: Allyn & Bacon.

Symonds, P. M. (1927). *Measurement in secondary education.* Nueva York: Macmillan.

Taylor, C. S., y Nolen, S. B. (2005). *Classroom assessment: Supporting teaching and learning in real classrooms.* Upper Saddle River, NJ: Pearson Education.

Teale, W. H., Hiebert, E. y Chittenden, E. (1987). Assessing young children's literacy development. *The Reading Teacher, 40,* 772-777.

Teale, W. H., y Sulzby, E. (Eds.). (1986). *Emergent literacy: Writing and reading.* Norwood, NJ: Ablex.

Tharp, R. G., y Gallimore, R. (1988). *Rousing minds to life: Teaching, learning, and schooling in social context.* Nueva York: Cambridge University Press.

Thorndike, E. L. (1913). *Introduction to the theory of mental and social measurements.* Nueva York: Teachers College, Universidad de Columbia.

Thorndike, E. L. (1922). Measurement in education. En *Twenty-first yearbook of the National Society for the Study of Education*. Parte 1, pp. 1-9. Bloomington, IL: Public School Publishing.

Thorndike, E. L. (1931). *Human learning.* Nueva York: Century.

Thorndike, E. L. (1971). *Educational measurement,* 2ª ed. Washington, DC: American Council on Education.

Torgerson, T. L., y Adams, G. S. (1954). *Measurement and evaluation for the elementary-school teacher with implications for corrective procedures.* Nueva York: Dryden Press.

Travers, R. M. W. (1955). *Educational measurement.* Nueva York: Macmillan.

Tyler, R. W. (1951). The functions of measurement in improving instruction. In E. F. Lindquist (Ed.), *Educational measurement,* pp. 47-67. Washington, DC: American Council on Education.

U. S. Congress, Office of Technology Assessment. (1992). *Testing in American schools: Asking the right questions* (OTA-SET-519). Washington, DC: U.S. Government Printing Office.

Valencia, S. W., y Calfee, R. C. (1991) The development and use of literacy portfolios for students, classes, and teachers. *Applied Measurement in Education, 4,* 333-346.

Van den Heuvel-Panhuizen, M. (2001). A learning-teaching trajectory description as a hold for mathematics teaching in primary schools in the Netherlands. En M. Tzekaki (Ed.), *Didactics of mathematics and informatics in education. 5ª Conferencia Panhelénica con Participación Internacional,* pp. 21-36. Thessaloniki: Universidad Aristóteles de Thessaloniki, Universidad de Macedonia, Instituto Pedagógico.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Ward, J. G. (1980). Teachers and testing: A survey of knowledge and attitudes. En L. M. Rudner (Ed.), *Testing in our schools*, pp. 15-24. Washington, DC: National Institute of Education.

White, B. Y., y Frederiksen, J. R. (2000). Metacognitive facilitation: An approach to making scientific inquiry accessible to all. En J. Minstrell y E. van Zee (Eds.), *Inquiring into inquiry learning and teaching in science,* pp. 33-370. Washington, DC: American Association for the Advancement of Science.

Wigheld, A., Eccles, J. S., y Rodríguez, D. (1998). The development of children's motivation

in school contexts. En P. D. Pearson y A. Iran-Nejad (Eds.), *Review of Research in Education, 23*, 73-118. Washington, DC: American Educational Research Association.

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership, 49*, 26-33.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan, 74,* 200-214.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance.* San Francisco: Jossey-Bass.

Wiggins, G., y McTighe, J. (1998). *Understanding by design.* Alexandria, VA: Association for Supervision and Curriculum Development.

Wood B. D. (1923). *Measurement in higher education.* Yonkerson-Hudson, Nueva York: Banco Mundial.

Wood, D., Bruner, J. S., y Ross, G. (1976). The role of tutoring in problem-solving. *Journal of Child Psychology and Psychiatry, 17*, 89-100.

Yeh, J. P., Herman, J. L., y Rudner, L. M. (1981). *Teachers and testing: A survey of test use* (CSE Report No. 166). Los Ángeles: UCLA, Center for the Study of Evaluation.