Archivos de Ciencias de la Educación, 2009 3(3). ISSN 2346-8866.
http://www.archivosdeciencias.fahce.unlp.edu.ar

Una síntesis integradora de la investigación e implicancias para una nueva teoría de la evaluación formativa*

*An integrative summary of the research literature and implications for a new theory of formative assessment**

Dylan Wiliam
Institute of Education, University of London

Si lo que los estudiantes aprendieron como resultado de una particular secuencia de instrucción fuera predecible, no habría necesidad de evaluación. Los evaluadores deberían simplemente compilar un inventario de lo que ellos han enseñado y usarlo como un catálogo de lo que los estudiantes han aprendido. Esta era, efectivamente, la asunción subyacente del modelo medieval de las universidades de Oxford y Cambridge, donde se entregaba un certificado de *Bachelor* después de completar un determinado período de residencia. Por supuesto que, tal como un vasto cuerpo de investigaciones (e.g., Denvir & Brown, 1986a, b) y la experiencia de cada educador indican, lo que los estudiantes aprenden de una particular secuencia de enseñanza puede ser muy diferente de lo que el docente pretendía. Es por eso que la evaluación es central y quizás un rasgo que define a la instrucción efectiva: la evaluación es la única manera de saber si lo que se ha enseñado fue aprendido. En un sentido muy real, por lo tanto, la evaluación es el puente entre la enseñanza y el aprendizaje.

Traducción de Estela B. Cols, Facultad de Humanidades y Ciencias de la Educación, UNLP. Supervisión Trad. Pilar Romero.

La evaluación es la que provoca la rutina de que los docentes y los estudiantes se unan con el propósito de crear una forma diferente de aprendizaje, por ejemplo, la de un docente hablando desde una video cámara que luego se transmite a los estudiantes que están en otra habitación: Juntos docentes y estudiantes pueden asegurarse de que la información acerca del rendimiento de los alumnos, obtenida a través de la evaluación, pueda ser usada para ajustar la instrucción en vistas a satisfacer mejor sus necesidades de aprendizaje. Esta es la esencia de la evaluación formativa: la idea de que la evidencia de los logros del estudiante es obtenida e interpretada, y conduce a una acción que resulta en un mejor aprendizaje que aquel que hubiera tenido lugar en la ausencia de tal evidencia.

Las descripciones acerca de los orígenes del término evaluación formativa pueden encontrarse en otros capítulos de este volumen y en Wiliam (2007a). El

Esta obra está bajo licencia Creative Commons Atribución-NoComercial-SinDerivadas 2.5 Argentina

propósito de este capítulo es partir de una idea básica de evaluación formativa para tratar de proveer una base teórica firme acerca de los modos en que ésta puede apoyar el aprendizaje, mostrar cómo las diversas formulaciones de la noción de evaluación formativa, que han sido propuestas en las últimos 40 años, pueden ser integradas dentro de un marco más amplio e indicar brevemente cómo este marco se conecta con las investigaciones en diferentes áreas.

Reseñas de la investigación sobre retroalimentación y evaluación formativa

Una de las más poderosas metáforas que subyacen a la teoría de la acción de la evaluación formativa es la idea de retroalimentación, desarrollada originalmente en el campo de los sistemas ingenieriles. Como señala Ramaprasad (1983), el rasgo distintivo de la retroalimentación es que la información generada dentro de un sistema debe tener algún efecto sobre éste. La información no tiene la capacidad de cambiar el desempeño del sistema sin retroalimentación: "La retroalimentación es información acerca de la distancia entre el nivel actual y el nivel de referencia de un parámetro del sistema que es utilizado para modificar esa brecha en alguna forma" (Ramaprasad 1983: 4). Comentando acerca de esto, Sadler señaló:

> "Un importante rasgo de la definición de Ramaprasad es que la información acerca de la distancia o brecha entre el nivel actual y el de referencia es considerada como retroalimentación *solo cuando es usada para alterar esa distancia.* Si la información solo es registrada, pasada a una tercera parte que no tiene ni el conocimiento ni el poder para cambiar el resultado, o está muy profundamente codificada (como por ejemplo, un resumen de calificaciones otorgado por un docente) para conducir a la acción apropiada, el circuito del control no puede cerrarse y la "información pendiente" no puede ser sustituida por retroalimentación efectiva" (1989:121).

Desde esta perspectiva, la retroalimentación no puede ser separada de sus consecuencias instruccionales. Por lo tanto, no es sorprendente que a lo largo del último cuarto de siglo, haya aparecido un número de sólidas reseñas acerca del impacto de las prácticas de evaluación en los estudiantes y en sus aprendizajes en el contexto del aula (Fuchs & Fuchs 1986; Natriello, 1987; Crooks, 1988; Bangert-Drowns, Kulik, Kulik & Morgan, 1991; Dempster, 1991, 1992; Elshout-Mohr, 1994; Kluger & DeNisi, 1996; Black & Wiliam, 1998a,b; Nyquist, 2003; Brookhart, 2004; Allal & Lopez, 2005; Köller, 2005; Brookhart, 2007; Wiliam, 2007a; Hattie & Timperley, 2007; Shute, 2008).

Una síntesis detallada de cada una de estas reseñas está fuera de los alcances de este capítulo, y además ellas se resisten a una síntesis fácil debido a las diferencias en los supuestos de los que parten, sus bases teóricas, y sus prescripciones (Brookhart, 2004). Sin embargo, algunos temas significativos emergen:

- El primero es que los resultados del análisis son usados en una multiplicidad de

formas, a menudo en conflicto (Natriello, 1987; Corooks, 1988; Black y Wiliam, 1998a). En particular el uso de evaluaciones para propósitos sumativos (como el de determinar la calificación en un curso) parece reducir el grado en el cual ella puede servir para apoyar el aprendizaje.

- El segundo es que diversos modos de retroalimentación pueden ser diferencialmente efectivos para diferentes tipos de aprendizajes. Por ejemplo, las formas de retroalimentación que son más efectivas en el desarrollo de habilidades de bajo nivel y conocimiento del contenido pueden no serlo tanto para las habilidades de orden superior (Demspster, 1991, 1992; Elshout-Mohr, 1994), y en particular, que la retroalimentación inmediata parece ser más efectiva para el aprendizaje procedural, mientras que la retroalimentación demorada puede serlo para las habilidades de orden superior (Shute, 2008).
- El tercero, y quizás más importante de los temas, es que la retroalimentación más efectiva focaliza su atención *prospectivamente* más que *retrospectivamente*. La pregunta importante no es, "¿Qué hice bien y qué hice mal?", sino "¿Qué hacer ahora?" (Fuchs & Fuchs, 1986; Bangert-Drowns et al., 1991; Nyquist, 2003; Hattie & Timperley, 2007). Los estudios de corto plazo suelen ser particularmente engañosos con respecto a este punto, porque si bien ciertos tipos de intervenciones retroalimentadoras -definidas como "acciones llevadas a cabo por (un) agente (s) externo para proveer información relativa a algún(os) aspecto(s) del propio desempeño en la tarea" (Kluger & DeNisi, 1996 p. 255)- pueden mejorar el desempeño, quizás logren hacerlo mediante un cambio en el tipo de motivación. Por ejemplo, una intervención de retroalimentación puede mostrar efectos positivos aumentando la motivación hacia la tarea, pero entonces, el aprendizaje futuro requerirá retroalimentación continua. Aún cuando el énfasis esté en los procesos de aprendizaje de tareas, las intervenciones de retroalimentación pueden promover aprendizajes superficiales, haciendo, en consecuencia, que las habilidades de nivel superior sean más difíciles de lograr (Kluger & DeNisi, 1996; Shute, 2008).

Magnitud del efecto en las reseñas de investigaciones acerca de la evaluación formativa y sus limitaciones

Las reseñas de investigaciones anteriormente citadas producen un rango de estimaciones acerca de la magnitud del efecto que puede esperarse que tenga el uso de la retroalimentación formativa en el aprendizaje. Bangert-Drowns et al. (1991) hallaron un efecto promedio de alrededor de un cuarto de desviación estándar para la retroalimentación en eventos "tipo test", mientras que Kluger y DeNisi (1996) y Nyquist (2003) encontraron que la retroalimentación producía efectos algo mayores -alrededor de 0.4 de desviación estándar (a pesar de que ambos advirtieron que la variabilidad a lo largo de diferentes estudios era extremadamente alta). Black y Wiliam (1998a) y Shute (2008) sugirieron que la medida típica de los efectos estaba en el rango de 0.4 a 0.7 y de 0.4 a 0.8 respectivamente mientras que una reseña de 74 meta-análisis

acerca de los efectos de la retroalimentación realizada por Hattie y Timperley (2007) encontró que el impacto promedio era de 0.95 desviaciones a través de 4.157 estudios.

El uso de medidas estandarizadas para comparar y sintetizar los estudios sobre los efectos es entendible, dado que pocos de los estudios incluidos en las distintas reseñas publican detalles suficientes como para permitir el desarrollo de formas de síntesis más sofisticadas, pero descansar en ellas crea dificultades importantes de interpretación cuando se trata de estudios educativos, por dos razones:

- La primera es que, como Black y Wiliam señalaron (1998a), la magnitud del efecto está influenciada por el rango de rendimiento de la población. Un incremento de cinco puntos en una prueba en la cual la desviación estándar de la población es de 10 puntos podría resultar en un efecto de 0,5 desviaciones estándar. Sin embargo, si se administra la misma intervención a la mitad superior de la misma población, asumiendo que fue igualmente efectiva para todos los estudiantes, podría resultar en un efecto de 0,8 desviaciones estándar, debido a la reducción de la dispersión en la sub-muestra. Un hallazgo frecuente en la literatura es que las intervenciones de evaluación formativa son más exitosas con los estudiantes que tienen necesidades educativas especiales (por ejemplo, en Fuchs & Fuchs, 1986 arriba). Pero esto resulta difícil de interpretar sin un intento de controlar la restricción del rango y puede ser simplemente un artilugio estadístico.
- La segunda, y más importante limitación de los meta-análisis discutidos anteriormente, es que ellos no logran considerar el hecho de que las diferentes medidas del resultado no son igualmente sensibles a la instrucción (Popham, 2007). Gran parte de la metodología de meta-análisis usada en educación y en psicología fue tomada acríticamente de las ciencias médicas y de la salud, en las que diferentes estudios combinados en un meta-análisis empleaban las mismas formas de medición de los logros (por ejemplo, tasas de supervivencia de un año) o medidas de logros que eran razonablemente consistentes a través de diferentes contextos (por ejemplo, el momento de alta de los cuidados hospitalarios). En educación, para agregar resultados provenientes de diferentes estudios es necesario asumir que las medidas del rendimiento son igualmente sensibles a la instrucción.

Hace tiempo que es sabido que las mediciones llevadas a cabo y construidas por los docentes tienden a mostrar efectos de mayor magnitud para las intervenciones experimentales que las obtenidas en pruebas estandarizadas y esto, algunas veces, fue considerado como una evidencia de la invalidez de las mediciones efectuadas por los docentes. Sin embargo, como ha quedado en claro en los últimos años, las evaluaciones varían ampliamente en su sensibilidad a la instrucción -el grado en el cual ellas miden las cosas que los procesos educativos cambian (Wiliam, 2007b). En particular, el modo en que las pruebas estandarizadas son construidas, reduce su sensibilidad a la enseñanza. La confiabilidad de una prueba puede ser aumentada reemplazando ítems que no permiten discriminar entre los distintos sujetos por ítems que sí lo hacen, de

modo que los ítems que todos los estudiantes responden correctamente y aquellos que todos los estudiantes responden incorrectamente, son generalmente omitidos. Sin embargo, esa supresión sistemática de ítems altera el constructo que la prueba está midiendo porque algunos aspectos relativos al aprendizaje que son efectivamente enseñados por los maestros tienen menos probabilidad de ser incluidos que otros temas que quizás no fueron bien enseñados.

Por ejemplo, un ítem que es respondido de modo incorrecto por todos los alumnos de 7º grado y es respondido bien por todos los alumnos de 8º grado, está seguramente evaluando algo que es modificado por la enseñanza (suponiendo que no se trata de algo trivial como "¿Cuál es el nombre de tu maestra de matemáticas de 8º año?") pero que tiene pocas probabilidades de ser mantenido en una prueba para alumnos de 7º grado (porque es demasiado difícil) ni tampoco en una de 8º (porque es demasiado sencillo). Es un ejemplo extremo, pero en verdad llama la atención acerca del modo en que la sensibilidad de una prueba a los efectos de la instrucción, puede verse afectada de modo significativo en el proceso normal de construcción de una prueba (Wiliam, 2008).

Los efectos de la sensibilidad a la instrucción están lejos de ser despreciables. Bloom (1984) observó que la enseñanza tutorial uno a uno era más efectiva que la enseñanza grupal por dos desviaciones estándar. Esta pretensión es creíble en el contexto de muchas evaluaciones pero para las pruebas estandarizadas como las utilizadas por la National Assessment of Educational Progress (NAEP), un progreso de un año para un estudiante promedio es equivalente a una cuarta parte de un desvío estándar (NAEP, 2006). Entonces para que la pretensión de Bloom sea cierta, un año de tutoría individual debería producir el mismo efecto que nueve años de la enseñanza grupal media, lo que parece improbable. El punto importante aquí es que las medidas del rendimiento usadas en diferentes estudios probablemente difieran en forma significativa en su sensibilidad a la instrucción, y el elemento más significativo para determinar esa sensibilidad a la instrucción parece ser la distancia con respecto al curriculum que se pretende evaluar.

Ruiz-Primo, Shavelson, Hamilton y Klein (2002) propusieron una clasificación de cinco grados para la distancia de una evaluación con relación al curriculum real, con ejemplos como los siguientes:

- *inmediata*, como los informes de ciencias, los cuadernos de clases, las pruebas de aula;
- *cercana* o evaluaciones "integradas formales" (por ejemplo, si la evaluación inmediata implicaba el número de oscilaciones del péndulo en 15 segundos, una evaluación cercana preguntaría por el tiempo que pueden tomar 10 oscilaciones);
- *proximal*, que incluye diferentes evaluaciones de un mismo concepto solicitando cierta transferencia (por ejemplo, si la evaluación inmediata solicitaba a los alumnos la construcción de pequeños botes a partir de vasos de papel, la evaluación próxima solicita una explicación de por qué los botes flotan o se hunden);

- *distal*, por ejemplo una evaluación a gran escala desde un marco de evaluación estatal, en el cual la consigna de evaluación fuera obtenida por muestreo a partir de un dominio diferente, como la ciencia física, y en la que el problema, los procedimientos, materiales y métodos de medición difieran de aquellos que fueron utilizados en las actividades originales;
- *remota*, como en el caso de las evaluaciones de desempeño estandarizadas a escala nacional.

Los autores mencionados encontraron, como podía esperarse, que cuanto más cercana es la evaluación al curriculum actuado, mayor es la sensibilidad de una evaluación a los efectos de la enseñanza, y que ese impacto era considerable. Por ejemplo, una de las intervenciones mostró un efecto promedio de 0.26 cuando fue medido con una evaluación proximal, mientras que ascendía a 1.26 al ser medido con una evaluación cercana.

En ninguno de los meta-análisis discutidos arriba hubo un intento de controlar los efectos de las diferencias de la sensibilidad a la enseñanza en las distintas medidas del logro. Esto no invalida de por sí la pretensión de que la evaluación formativa permite mejorar los logros de los estudiantes. En verdad, con toda probabilidad, los intentos de mejorar la calidad de las prácticas de evaluación formativa de los docentes tienen probabilidades de ser más rentables, sino las más, que muchas otras intervenciones (Wiliam & Thomson, 2007). Sin embargo, la falta de control del impacto de este factor significa que hay que tener un cuidado considerable en citar valores particulares del efecto como si fueran posibles de ser alcanzados en la práctica, y que otras medidas del impacto, como el incremento en las tasas de aprendizaje, pueden ser más apropiadas (Wiliam, 2007c). Más importante aún es que necesitamos mover nuestra atención desde el tamaño del efecto hacia el rol que puede jugar la retroalimentación en el diseño de contexto de aprendizajes efectivos. (Wiliam, 2007a). Al concluir su reseña acerca de 3000 estudios sobre el impacto de las intervenciones de retroalimentación en escuelas, colleges y lugares de trabajo, Kluger y DeNisi observaron:

> "Las consideraciones acerca de las intervenciones alternativas y de utilidad sugieren que aun una intervención retroalimentadora con efectos positivos demostrados, no debe ser administrada dondequiera que sea posible. Más bien es necesario un desarrollo mayor de la FIT [Teoría de intervención retroalimentadora/feedback intervention theory] para establecer las circunstancias bajo las cuales los efectos positivos de la intervención retroalimentadora sobre el desempeño son duraderos y a la vez eficientes y cuándo esos efectos son transitorios y de utilidad dudosa. Esta investigación debe focalizarse en el proceso inducido por la intervención retroalimentadora y no sobre una pregunta general acerca de si la intervención retroalimentadora mejora el desempeño -observen a qué pocos progresos han conducido 90 años de intentos de responder a esta pregunta" (1996:278).

En el resto de este capítulo se analizan una serie de definiciones recientes de la evaluación formativa y se propone una definición en términos de la función que cumple la evidencia obtenida a través de la evaluación, específicamente la medida en que la misma apoya y mejora las decisiones instruccionales. Se examinan luego las consecuencias de esta definición, concentrándose en particular en el modo en que la evaluación formativa puede ponerse en práctica, y el capítulo concluye esbozando brevemente algunos vínculos con otras áreas de investigación afines y algunas prioridades para futuras investigaciones.

Definiciones de evaluación formativa

A lo largo de los años se han propuesto una variedad de definiciones del término "evaluación formativa". En su reseña, Black y Wiliam definieron la evaluación formativa como "aquella que abarca todas las actividades llevadas a cabo por los docentes, y/o por sus estudiantes, las cuales proveen información para ser usada como retroalimentación para modificar las actividades de enseñanza y de aprendizaje en las que están involucrados" (1998ª: 7). En una publicación posterior, dirigida a los encargados de la definición de políticas y a los docentes, adoptaron la siguiente definición:

> "Usamos el término general *evaluación* para referirnos a todas aquellas actividades llevadas a cabo por los docentes -y por los estudiantes cuando se evalúan a sí mismos- que proveen información para ser usada como retroalimentación para modificar las actividades de enseñanza y aprendizaje. Esa evaluación se vuelve *evaluación formativa* cuando la evidencia es efectivamente usada para adaptar la enseñanza a las necesidades de los alumnos" (Black & Wiliam, 1998b: 140).

Cowie y Bell (1999) adoptaron una definición ligeramente más restrictiva al limitar el término a la evaluación conducida *y* actuada, mientras el aprendizaje está teniendo lugar. Los autores definieron la evaluación formativa como "el proceso usado por los docentes y estudiantes para reconocer y responder al aprendizaje de los alumnos en orden a mejorar ese aprendizaje, durante el aprendizaje"(Cowie y Bell, 1999: 32). El requisito de que la evaluación sea llevada a cabo durante el aprendizaje, también fue propuesto por Shepard, Hammerness, Darling-Hammond y Rust en su definición de la evaluación formativa como "la evaluación llevada a cabo durante el proceso instructivo con el propósito de mejorar la enseñanza o el aprendizaje" (2005: 275). En la reseña de prácticas de evaluación formativa desarrolladas en ocho sistemas educativos nacionales y provinciales, la OCDE también enfatizó el principio de que la evaluación debe tener lugar durante la instrucción: "La evaluación formativa refiere a la evaluación frecuente e interactiva del progreso y la comprensión de los estudiantes con el fin de identificar las necesidades de los estudiantes y ajustar apropiadamente la enseñanza" (Looney, 2005: 21). En una línea similar, Kahl escribió: "La evaluación

formativa es una herramienta que los docentes usan para medir la captación por parte de los estudiantes de los temas específicos y habilidades que están enseñando. Es una herramienta que está "a mitad de la corriente" y permite identificar los errores y concepciones alternativas específicas de los alumnos mientras el material está siendo enseñado" (2005: 11).

Broadfoot, Daugherty, Gardner, Gipps, Harlen, James y Stobart (1999) sostuvieron que la mejora del aprendizaje a través de la evaluación formativa dependía de cinco factores claves: 1) la provisión de retroalimentación efectiva a los alumnos, 2) el involucramiento activo de los alumnos en su propio aprendizaje, 3) el ajuste de la enseñanza teniendo en cuenta los resultados de la evaluación, 4) un reconocimiento de la profunda influencia que tiene la evaluación sobre la motivación y la autoestima de los alumnos, las cuales tienen cruciales influencias en el aprendizaje; y 5) la necesidad de los alumnos de ser capaces de evaluarse a sí mismos y comprender cómo mejorar. Los autores sugirieron que el término evaluación formativa no era de utilidad para describir esos usos de la evaluación "porque el término 'formativa' en sí mismo está abierto a una variedad de interpretaciones y a menudo no significa otra cosa que la evaluación es llevada a cabo frecuentemente y es planificada al mismo tiempo que la enseñanza" (Broadfoot, Daugherty, Gardner, Gipps, Harlen, James y Stobart, 1999: 7) A cambio, propusieron el uso del término "evaluación para el aprendizaje".

El primer empleo del término "evaluación para el aprendizaje" parece estar en una presentación efectuada en la conferencia anual de la Association for Supervision and Curriculum Development (James, 1992); el mismo año fue publicado en un libro titulado "Testing for learning" (Mitchell, 1992). Y el término "evaluación para el aprendizaje" fue utilizado como título de un libro tres años después (Sutton, 1995) pero el primer uso del término evaluación *para* el aprendizaje como opuesto a evaluación *del* aprendizaje parece haber sido realizado por Gipps y Stobart (1997). El uso del término se hizo popular en el Reino Unido gracias a Broadfoot et al. (1999) y en los Estados Unidos a Stiggins (2002).

La definición dada por el *Assessment Reform Group* [Grupo para la Reforma de la evaluación] (Broadfoot, Daugherty, Gardner, Harlen, James & Stobart) es: "la evaluación para el aprendizaje es el proceso de búsqueda e interpretación de evidencias para ser usada por los estudiantes y sus docentes para decidir dónde se encuentran los aprendices en sus procesos de aprendizaje, hacia dónde necesitan dirigirse y cuál es el mejor modo de llegar hasta allí" (2002: 2-3).

Mientras varios autores han utilizado los términos evaluación formativa y evaluación para el aprendizaje de modo indistinto, o como diferentes modos de rotular la misma idea, Black, Harrison, Lee, Marshall y Wiliam marcaron una distinción entre ambos términos:

> "Evaluación para el aprendizaje es cualquier evaluación cuya principal prioridad en su diseño y su puesta en práctica es la de servir al propósito de promover el aprendizaje del alumno. Difiere de la evaluación diseñada prin-

> cipalmente para servir al propósito de la acreditación, a del establecimiento de rankings o de certificar competencias. Una actividad de evaluación puede ayudar al aprendizaje si provee información que los docentes y sus estudiantes pueden usar como retroalimentación al evaluarse a sí mismos o a otros y al modificar las actividades de enseñanza y aprendizaje en las que están implicados. Esa evaluación se vuelve evaluación formativa cuando la evidencia es efectivamente usada para adaptar la tarea de enseñanza a las necesidades del aprendizaje" (2004: 10).

Quizás el punto más importante es la distinción entre evaluación formativa y evaluación sumativa en términos de la función que cumple la evaluación, más que la evaluación en sí. Wiliam y Black (1996) argumentaron que el intentar usar las palabras formativa y sumativa para describir a la evaluación lleva a una contradicción, dado que el mismo instrumento de evaluación, y aun los mismos resultados de la evaluación, pueden ser usados formativa y sumativamente. El hecho de localizar la distinción en términos del propósito de la evaluación permite salvar algunas de estas dificultades. Queda todavía abierta la posibilidad de que la evidencia evaluativa pueda ser recogida con la intención de apoyar el aprendizaje, pero que tal vez ello nunca ocurra.

Una nueva teoría de la evaluación formativa: precisión en la definición

Con el fin de ofrecer una definición comprensiva de evaluación formativa, Black y Wiliam propusieron la siguiente:

> "La práctica en una clase es formativa en la medida en que la evidencia acerca de los logros de los estudiantes es obtenida, interpretada y usada por docentes, aprendices o sus pares para tomar decisiones acerca de sus próximos pasos en la instrucción que tengan probabilidades de ser mejores, o mejor fundadas, que las decisiones que ellos hubieran tomado en la ausencia de la evidencia que fue obtenida" (2009:.6).

Para explicar esta definición, Black y Wiliam (2009) señalan los siguientes puntos:

1. *Cualquiera puede ser el agente de la evaluación formativa.* Mientras en muchos casos las decisiones podrán ser tomadas por el docente, la definición también incluye aquellas situaciones en las cuales las decisiones son tomadas por los propios aprendices, o por sus pares.

2. *El foco de la definición está en las decisiones.* Black y Wiliam señalan que el foco de la definición podría estar en las intenciones de aquellos involucrados en la instrucción de recoger la evidencia, pero entonces las actividades de recolección de datos que no tuvieron impacto alguno en el aprendizaje podrían ser potencialmente formativas, lo cual sería contrario al sentido común (y, en verdad, al significado literal del término formativo). Tal definición *sería*, en tal sentido, demasiado abierta. Por otro lado, la definición de Black y Wiliam (1998b) pone el foco en el resultado. Re-

quiere que la evaluación en efecto conduzca a un mejor aprendizaje, lo cual parecería ser un criterio un tanto restrictivo, en tanto podría haber muchas situaciones en las cuales las acciones de las que podría esperarse que mejoren el aprendizaje, podrían no hacerlo dada la impredecible naturaleza del aprendizaje (¡y de los estudiantes!). El foco en las decisiones es también consistente con la definición de pedagogía de Robin Alexander como:

> "El acto de enseñar juntos con su correspondiente discurso de teorías, valores, evidencia y justificaciones. Es lo que uno necesita saber, y las habilidades que uno necesita dominar, en vistas a tomar y justificar los muchos diferentes tipos de decisiones de los que el aprendizaje está constituido" (2008: 47)

3. *La definición pone el foco en los próximos pasos en la instrucción.* El término "instrucción" es usado para describir cualquier actividad intencional para crear el aprendizaje, que aquí es definido como un incremento, derivado de la experiencia, en las capacidades de un individuo de actuar o reaccionar frente a los estímulos, en unas formas valiosas. Por lo tanto, el término instrucción subsume los roles del docente y del aprendiz. Este empleo del término será poco familiar para algunos lectores dado que es utilizado en determinados contextos para denotar un enfoque de enseñanza "transmisivo" pero de ninguna manera ésta es la connotación que se pretende aquí. En este marco resulta valioso señalar que existen algunos idiomas en los cuales se utiliza la misma palabra para designar a la enseñanza y al aprendizaje (Galés: *dysgu*; Maori: *ako*). Es este sentido inclusivo de la palabra instrucción, denotando a la enseñanza y al aprendizaje, a la que se alude aquí.

4. *La definición es probabilística.* Al localizar el peso de la definición del término "formativa" en la acción resultante crea la dificultad de que es imposible establecer la prueba del efecto, requiriendo la verificación de una demanda contra-fáctica: que lo que ocurrió fuera diferente (o mejor que) lo que hubiera sucedido en la ausencia de esa evaluación (pero que no sucedió). El hecho de requerir que las decisiones tengan probabilidades de ser mejores refleja el hecho de que aun las intervenciones mejor diseñadas pueden no resultar *siempre* en mejor aprendizaje para *todos* los estudiantes.

5. *La evaluación no necesita cambiar la instrucción planificada.* La definición requiere que las decisiones sean o bien mejores o mejor fundadas que las decisiones tomadas sin la evidencia recogida como parte del proceso de evaluación. La segunda posibilidad es incluida para contemplar aquellos casos en los que la evaluación indica al docente que el mejor curso de acción es en verdad aquel que había intentado antes de la obtención de la evidencia. En este caso, la evaluación formativa no cambiará el curso de acción pero significará que está mejor fundada en la evidencia. (Para este punto agradezco a Jim Popham, quien a través de una implacable exploración, obligó a una clarificación de este aspecto de la definición).

A partir de esta definición, Black y Wiliam proponen que la evaluación formativa es, en esencia, concerniente a "la creación de y capitalización de `momentos de

contingencia´ en la instrucción con el propósito de regular los procesos de aprendizaje" (2009: 6) Una teoría de la evaluación formativa es, por lo tanto, mucho más estrecha que una teoría integral de la enseñanza y el aprendizaje, a pesar de que se liga de un modo significativo a otros aspectos de la enseñanza y el aprendizaje. Dado que la manera en que los docentes, los aprendices y sus pares crean y capitalizan esos momentos de contingencia implica consideraciones de diseño instruccional, curriculum, pedagogía, psicología y epistemología.

Los momentos de contingencia pueden ser sincrónicos o asincrónicos. Los ejemplos de momentos sincrónicos incluyen los ajustes que los maestros realizan en tiempo real durante una enseñanza uno a uno o en una discusión con toda la clase. Los ejemplos asincrónicos incluyen la retroalimentación del docente, el uso de evidencia derivada de la tarea fuera de la escuela o las síntesis efectuadas al terminar una clase (por ej.: los "permisos de salida"*), usadas para planificar una clase posterior. Además, estos momentos asincrónicos podrían ser usados para modificar la instrucción acerca de la cual la evidencia fue recogida, o el docente puede recoger evidencia acerca de las dificultades experimentadas por un grupo, y utiliza esto para cambiar la instrucción de otro grupo de estudiantes en algún momento en el futuro.

Las respuestas de los profesores a la información acerca del aprendizaje de los estudiantes puede ser uno a uno o basada en el grupo. Las respuestas a un trabajo escrito es usualmente uno a uno pero en las discusiones en clase la retroalimentación está en relación con las necesidades de la clase como un todo y puede ser una intervención inmediata en el flujo de la discusión o una decisión acerca de cómo comenzar la próxima clase.

Una nueva teoría de la evaluación formativa: consecuencias de la definición

En esta sección se exploran dos consecuencias particulares de la definición de evaluación formativa propuesta arriba: los tipos de decisiones que las evaluaciones formativas pueden apoyar, y la inmediatez de los ajustes instruccionales que son informados por las evaluaciones.

¿Qué tipo de evaluaciones son formativas?

De la definición de evaluación formativa propuesta se desprende que toda evaluación que provee evidencia que tiene el potencial de mejorar la toma de decisiones instruccionales puede ser formativa, tanto si estas decisiones son tomadas por los docentes, los pares o los aprendices por sí mismos. La evaluación puede solamente *monitorear* los logros de los alumnos indicando que para ciertos estudiantes la instrucción no tuvo éxito. Si el docente entonces organiza instrucción adicional para esos estudiantes, aun si eso es volver sobre el material de modo más lento, entonces esto es potencialmente formativo. Si la evaluación provee información adicional que identifica la naturaleza precisa de las dificultades de los alumnos, entonces es *diagnóstica.* Las evaluaciones más provechosas, sin embargo, son aquellas que dan lugar a interpreta-

ciones que son *instruccionalmente tratables*. En otras palabras, no sólo ellas permiten identificar qué estudiantes están teniendo dificultades (la evaluación de monitoreo) o localizar la especificidad de esas dificultades (la evaluación diagnóstica): ellas generan comprensiones acerca de los próximos pasos de la instrucción (incluyendo posibles pasos a tomar por los alumnos) que tienen probabilidad de ser más efectivos.

Para tomar un ejemplo concreto, supongamos que una clase ha tomado una prueba que evalúa la capacidad de identificar la fracción más grande y la más pequeña en un conjunto dado. El conocimiento de los resultados alcanzados por los estudiantes en la prueba podría proveer evaluación de *monitoreo*. Permitiría identificar aquellos estudiantes que dominan suficientemente esa habilidad y que pueden seguir adelante, y los que necesitan más ayuda. Si el docente organizara instrucción adicional para estos últimos estudiantes, proponiendo una clase complementaria al final del día, o a través de la provisión de materiales orientados al logro de ese aprendizaje, la prueba podría ser formativa (o más precisamente, podría funcionar formativamente), porque la disponibilidad de los resultados de la prueba permitió al docente tomar mejores decisiones de las que hubiera podido tomar en ausencia de esa información.

Si la prueba ha sido cuidadosamente construida, puede haber además información *diagnóstica* en las respuestas de los estudiantes. Por ejemplo, el docente puede advertir que la mayoría de los estudiantes que obtuvieron puntajes bajos en la prueba tuvieron un desempeño mejor en ítems que incluían una cantidad de fracciones unitarias (cuyo numerador es 1) que en ítems sin ese tipo de fracciones. Si bien esa podría ser información de utilidad, esa comprensión permite más bien localizar la dificultad en el aprendizaje que indicar qué debe hacerse para superarla -el docente puede focalizar su intervención en fracciones no unitarias, lo cual parece ser más apropiado que volver a enseñar el tema en forma completa. Sin embargo, si el docente puede advertir a partir de las respuestas que los estudiantes están operando con una estrategia *naif* que sostiene que la menor fracción es aquella que tiene el denominador mayor, y que la mayor fracción es aquella que posee el denominador más pequeño -una estrategia que es exitosa con fracciones unitarias (Vinner, 1997)- entonces esto provee información para el docente que es *instruccionalmente tratable*. Estas evaluaciones *señalan* el problema (monitoreo) y *lo localizan* (diagnóstico). Ellas sitúan el problema en el marco de una teoría de la acción que permite sugerir medidas que pueden ser tomadas para mejorar el aprendizaje. La mejor evaluación formativa es entonces aquella que identifica recetas para la acción futura.

Adviértase que en estos tres escenarios relativos al ítem de fracciones, la evaluación funcionó formativamente en cada uno de los casos, dado que la información fue usada para tomar decisiones que tenían probabilidades de ser mejores que aquellas que se hubieran tomado en la ausencia de esa evidencia. Sin embargo, el hecho de que en los tres casos la evaluación haya funcionado formativamente no significó que cualquiera de los tres modos de usar la evidencia tuviera las mismas posibilidades de ser

efectivo. Por definición, las evaluaciones que dan lugar a interpretaciones diagnósticas tienen mayores probabilidades de conducir a mejores decisiones instruccionales que las que simplemente monitorean el logro de los estudiantes, y aquellas que posibilitan interpretaciones, que son *instruccionalmente tratables,* son aún mejores.

Una de las diferencias entre las evaluaciones que monitorean de las que diagnostican y de las que dan lugar a interpretaciones que son instruccionalmente tratables, es un tema de especificidad de la información recogida. Para ser instruccionalmente tratable, la evaluación necesita recoger más información que simplemente si el aprendizaje tiene lugar o, de no ser así, qué es específicamente lo no aprendido. Para que el resultado de una evaluación sea instruccionalmente tratable debe involucrar teorías del curriculum y del aprendizaje.

La "tratabilidad instruccional" involucra una teoría del curriculum porque el foco está en "¿ahora qué sigue?" implicando que hay una clara noción de una progresión de aprendizajes: una descripción del "conocimiento, las habilidades, las comprensiones, las actitudes o valores que los estudiantes desarrollan en un área de aprendizaje, en el orden en el cual ellos típicamente los desarrollan" (Forster & Masters, 2004: 65). La tratabilidad instruccional también involucra una teoría del aprendizaje, porque antes que se pueda tomar una decisión acerca de qué tipo de evidencia obtener, es necesario conocer no sólo qué es lo que sigue en términos de aprendizaje, sino también qué tipo de dificultades tienen los aprendices al atravesar esos pasos. Las relaciones entre evaluación formativa y teorías del aprendizaje están más desarrolladas en Black y Wiliam (2005), Brookhart (2007), Wiliam (2007a) y en Black y Wiliam (2009), y serán brevemente sintetizadas en la sección siguiente sobre procesos instruccionales claves.

La longitud de los ciclos de evaluación formativa

En el ejemplo de las fracciones discutido arriba, la acción tomada por el docente sigue rápidamente luego de la obtención de la evidencia acerca del desempeño de los estudiantes. Sin embargo, en general, la evaluación formativa permite ciclos de obtención, interpretación y acción de diversa duración, asumiendo que la información es usada para dar cuenta de decisiones instruccionales. Consideren los siguientes seis ejemplos:

1. En la primavera de 2005, una supervisora del área de ciencias en una escuela del distrito necesita planificar talleres de verano que serán ofrecidos a los docentes de ciencias de 8º año del distrito. Ella analiza los puntajes obtenidos por los estudiantes de 8º año en las evaluaciones de 2004 y advierte que, mientras los puntajes promedio de ciencias son comparables con los puntajes provinciales, el desempeño en ítems acerca de la Tierra son mucho más bajos que los puntajes promedio de la jurisdicción. Decide entonces poner el foco en las Ciencias de la Tierra. en las actividades profesionales a desarrollar durante el verano de 2005. Los talleres tienen una alta concurrencia por parte de los docentes de ciencias de 8º año. Los docentes vuelven a clase en octubre de 2005 e implementan sus métodos revisados basados en lo que aprendieron durante

el verano. Como resultado de ello, el desempeño de los estudiantes de 8º en ítems relativos a las Ciencias de la Tierra mejoró en las pruebas tomadas en el 2006.

2. Cada año un grupo de docentes de álgebra del *High school* revisa el desempeño de los estudiantes en una prueba estatal de Álgebra I. Ellos observan el grado de facilidad de cada ítem (la proporción de respuestas correctas). En los casos en los que la facilidad del ítem es más baja que la esperada, miran cómo fue planificada y llevada a cabo la instrucción de ese aspecto del curriculum, y los modos en que la enseñanza puede ser fortalecida el año siguiente.

3. Un distrito escolar usa una serie de pruebas interinas que están basadas en el curriculum y son administradas cada 6 a 10 semanas para controlar el progreso de los estudiantes. Aquellos cuyos puntajes están por debajo del umbral determinado como necesario para tener un 80% de posibilidades de aprobar la prueba estatal, son obligados a concurrir a una instrucción adicional los días sábados.

4. En la enseñanza de ciencias y matemáticas en la escuela básica y media en Japón, una unidad de enseñanza está compuesta por 13 o 14 lecciones (Lewis, 2002). El contenido usualmente ocupa 10 u 11 lecciones, permitiendo que en la clase 11 ó 12 se administre una breve prueba, y que el docente utilice las lecciones restantes para volver a enseñar aspectos de la unidad que no fueron bien comprendidos.

5. Durante los últimos tres minutos de su clase, un profesor de historia que ha estado enseñando acerca de los problemas del sesgo histórico solicita a sus alumnos que respondan la siguiente pregunta en una ficha de 3 por 5 pulgadas: "¿Por qué los historiadores están preocupados acerca del sesgo histórico en las fuentes históricas?" Los estudiantes entregan estos "permisos de salida" cuando abandonan la clase. El profesor lee las respuestas dadas y luego tira los "permisos de salida" tras decidir que las respuestas de los estudiantes indican una comprensión suficientemente buena como para proseguir con la enseñanza de un nuevo capítulo en la clase siguiente.

6. Una profesora de ciencias de la escuela media ha estado enseñando a los estudiantes acerca de los diferentes tipos de palancas. Después de haber explicado que el principio clave de la clasificación concierne al particular arreglo de la carga, el esfuerzo y el punto de apoyo, ilustra este principio con tres ejemplos: un balancín (tipo 1), una carretilla (tipo 2) y una caña para pescar en mar profundo (tipo 3). Para comprobar la comprensión de los estudiantes, pregunta a la clase cómo podrían clasificarse un par de pinzas, pidiendo a cada estudiante que levante uno, dos o tres dedos para indicar su respuesta. Ella se sorprende de que la mayoría de los estudiantes consideran que las pinzas corresponden al tipo 2. Cuando les pregunta por qué, los alumnos responden que esto se debe a que hay dos brazos para las pinzas. La profesora advierte que los estudiantes necesitan comprender que es la particular distribución de la carga, el esfuerzo y el punto de apoyo lo que es importante y no la cantidad de componentes, mediante la presentación de más ejemplos, como un par de tijeras y de un rompenueces.

Recordemos la definición de evaluación formativa propuesta por Black y Wiliam:

> "La práctica en una clase es formativa en la medida en que la evidencia acerca de los logros de los estudiantes es obtenida, interpretada y usada por docentes, aprendices o sus pares para tomar decisiones acerca de sus próximos pasos en la instrucción que tengan probabilidades de ser mejores, o mejor fundadas, que las decisiones que ellos hubieran tomado en la ausencia de la evidencia que fue obtenida" (2009: 6).

De acuerdo con esta definición, en cada uno de los seis ejemplos la evaluación funcionó formativamente porque la evidencia obtenida en la evaluación fue interpretada y utilizada para tomar decisiones que tendrían probabilidad de ser mejores (o en el caso del ejemplo 5, mejor fundadas) que aquellas decisiones que se podrían haber tomado en ausencia de tal evidencia. La duración del ciclo de evaluación formativa fue además ajustada a la capacidad del sistema para responder a la evidencia generada -por ejemplo, tiene poco sentido generar información diariamente si las decisiones a las que esa evidencia debe informar son tomadas solo mensualmente (Wiliam y Thompson, 2007).

Sin embargo, varios de estos seis ejemplos no podrían ser considerados formativos según algunas de las definiciones consideradas anteriormente. En particular, Shepard (2007) y Kahl (2005) se resistirían a considerar que el uso de la evaluación en los ejemplos uno, dos y tres, es formativo. Ellos señalan con razón que muchos diseñadores de pruebas adoptaron acríticamente el rótulo "formativo" y a menudo lo han aplicado simplemente a pruebas originalmente destinadas a cumplir una función sumativa (ver también Popham, 2006). Shepard señala que "lo que hace a una evaluación *formativa* es que es inmediatamente usada para efectuar ajustes para *dar forma* al nuevo aprendizaje" (2007: 281). Entonces, en cada uno de los seis ejemplos, la evidencia de la evaluación fue utilizada a fin de efectuar ajustes para dar forma al nuevo aprendizaje. Los ejemplos uno, dos y tres no llegan a cumplir el requisito de la inmediatez impuesto por Cowie y Bell (1999), Looney (2005) y Shepard (2007), pero posiblemente, lo mismo sucede con el ejemplo cuatro, dependiendo de la definición de inmediatez de cada uno.

La literatura proveniente de la investigación ofrece soporte a la aseveración de que los ejemplos cuatro, cinco y seis tienen más probabilidades de mejorar el aprendizaje y en mayor medida, que los usos en los ejemplos uno, dos y tres. En verdad, Shepard (2007) argumenta que hay relativamente poca evidencia de que las intervenciones como las de los ejemplos uno, dos y tres tengan probabilidades de tener impacto alguno. Sin embargo, parece demasiado extraño decir que estos ejemplos no son formativos con el fin de reservar el término formativo para aquellos tipos de evaluación que hacen una verdadera diferencia en los logros de los alumnos. En cambio, parece tener más sentido -y violentar menos el uso vernáculo de la palabra- decir que en los casos en los que la evaluación direcciona el futuro aprendizaje puede ser descrita como formativa, pero reconociendo que hay diferentes tipos de ciclos de duración en la evaluación

formativa, como lo propusieron Wiliam y Thompson (2007), y que pueden verse en la tabla 1.

Tabla 1: Duración de los ciclos para la evaluación formativa

Tipo	Enfoque	Duración
Ciclo largo	A través de los períodos señalados, cuartos, semestres, años	4 semanas a 1 año
Ciclo mediano	Durante y entre unidades instruccionales	1 a 4 semanas
Ciclo corto	Durante y entre lecciones	Día a día: 24 a 28 horas Minuto a minuto: 5 segundos a 2 horas

También es, posiblemente, una buena *realpolitik* cuando parece poco probable que los editores de pruebas concuerden en renunciar a las ventas adicionales que esperan lograr si marcan dichas pruebas como "formativas" (y por lo tanto reivindicar a todo un cuerpo de investigaciones sobre la eficacia en la práctica) simplemente porque se lo piden los investigadores. La pregunta importante no es, por lo tanto: ¿es esta evaluación formativa? Sino ¿de qué modo el uso de esta evaluación mejora el aprendizaje? Y haciéndonos eco de las conclusiones de Kluger y De Nisi (1996), ¿de qué modo sustentable esta evaluación mejora el aprendizaje?

Para responder a esta última pregunta y para entender qué modos de evaluación formativa tiene más probabilidades de ser efectiva, es necesario ir más allá de la definición funcional y mirar con mayor detalle los procesos que subyacen.

Una nueva teoría de la evaluación formativa: procesos instruccionales claves

El enfoque de sistemas de la evaluación formativa propuesto por Ramaprasad (1983) y que provee la base de la definición de "evaluación para el aprendizaje" adoptada por el *Assessment Reform Group* (Broadfoot y otros, 2002) pone la atención en tres procesos instruccionales:

1. Establecer dónde están los alumnos en sus procesos de aprendizaje.
2. Establecer hacia dónde están yendo.
3. Establecer qué es necesario hacer para que ellos lleguen allí.

La definición de evaluación formativa adoptada aquí está basada en el cruce de la dimensión de proceso (dónde están los alumnos en su proceso de aprendizaje, hacia dónde van y cómo llegar hasta allí) con el agente del proceso instruccional (docente, par, aprendiz). Las nueve celdas resultantes pueden reducirse a cinco estrategias claves de evaluación formativa tal como se muestra en la Figura 1. El foco de la Figura 1 es el contenido de la clase. Como Black y Wiliam observaron, las actividades que tienen

lugar cuando los estudiantes están aprendiendo Matemáticas son muy diferentes de aquellas que tienen lugar cuando están aprendiendo Lengua o Artes. El rol de los estudiantes y del docente, y la naturaleza de las interacciones entre sí y con la disciplina serán probablemente diferentes también. Además, el contenido de la clase está, por supuesto, anidado en una escuela que, a su vez está localizada en una comunidad, y así sucesivamente. Cualquier consideración acerca de la evaluación formativa debe poder reconocer esos múltiples contextos pero ellos están fuera del alcance de este capítulo. Desde que la posición tomada en este capítulo es que, en definitiva, la evaluación debe alimentar a las acciones relativas al contenido de la clase, en vistas a impactar en el aprendizaje, esta simplificación parece razonable, al menos en una aproximación de primer orden (para ejemplos de enfoques socioculturales de la implementación de la evaluación formativa, ver Black y Wiliam (2005) y Pryor y Crossouard (2005)).

Figura 1: Aspectos de la evaluación formativa

(Los números entre paréntesis indican a cuál de las cinco estrategias claves está relacionado el proceso)

<table>
<tr><th></th><th>Dónde va el aprendiz</th><th>Dónde está ahora el aprendiz</th><th>Cómo llegar ahí</th></tr>
<tr><td>Profesor</td><td>Clarificar y compartir las metas del aprendizaje y los criterios de logro (1)</td><td>Diseñar discusiones de clase efectivas, preguntas y tareas que permitan obtener evidencias del aprendizaje (2)</td><td>Proveer retroalimentación que permite a los alumnos avanzar (3)</td></tr>
<tr><td>Compañero</td><td>Comprender y compartir las metas del aprendizaje y los criterios de logro (1)</td><td colspan="2">Promover a los estudiantes como recursos de enseñanza para otros alumnos (4)</td></tr>
<tr><td>Aprendiz</td><td>Comprender las metas del aprendizaje y los criterios de logro (1)</td><td colspan="2">Promover a los estudiantes como los dueños de su propio aprendizaje (5)</td></tr>
</table>

El marco representado en la Figura 1 sugiere que la "evaluación para el aprendizaje" puede ser conceptualizada como aquella constituida por cinco estrategias claves (Wiliam & Thompson, 2007):

1. clarificar, compartir y comprender las metas de aprendizaje y los criterios de logro;
2. diseñar discusiones de clase efectivas, preguntas y tareas que permitan obtener evidencias acerca del aprendizaje;
3. proveer retroalimentación que permita a los alumnos avanzar;
4. promover a los estudiantes como recursos de enseñanza para otros alumnos;

5. promover a los estudiantes como los dueños de su propio aprendizaje.

En Wiliam (2007a) puede encontrarse un informe más detallado de estas estrategias. En lo que queda del capítulo sintetizaremos brevemente cada una de estas estrategias y concluiremos con algunas ideas acerca de las direcciones futuras de investigación, teoría y práctica.

Clarificar, compartir y comprender las metas de aprendizaje y los criterios de logro

La primera estrategia implica clarificar, comunicar y comprender las metas de aprendizaje y los criterios de éxito con los estudiantes. En ciertas ocasiones será posible especificar las intenciones en términos de metas claras, con criterios de logro estrechamente derivados de ellas, por ejemplo, cuando el docente está tratando de ayudar a los estudiantes a aprender cómo balancear una ecuación química. Otras veces, particularmente en el trabajo creativo, esa precisión no será posible ni deseable, por ejemplo, cuando los estudiantes están implicados en explorar las posibilidades que ofrece la pintura con acrílicos. En esas situaciones, el docente puede estar operando con un "horizonte" amplio (Black et al, 2003: 68) de metas posibles y aceptables; diversos estudiantes pueden seguir por diferentes caminos. Sin embargo, es importante notar que no se trata de que "todo vale". Si bien puede haber un amplio rango de diferentes direcciones por las que los alumnos pueden ir de modo sustantivo, debe haber algunas que el docente considera con pocas probabilidades de conducir a un aprendizaje provechoso, punto en el cual el docente probablemente intervendrá para redireccionar las actividades del aprendiz.

Una consecuencia importante de esta visión de la evaluación formativa es que, mientras es necesario clarificar qué es lo que debe ser aprendido, lo que los alumnos aprenderán es completamente independiente de la evaluación formativa (Wiliam, 2007a). En otras palabras, un compromiso hacia la evaluación formativa no implica ninguna visión particular acerca de cuáles deben ser las intenciones pedagógicas, ni tampoco involucra un compromiso especial con una visión acerca de lo que sucede cuando el aprendizaje tiene lugar. Esto es importante porque, en muchas formulaciones de evaluación formativa, hay un supuesto de que un compromiso con la evaluación formativa implica también un compromiso con ciertos tipos de objetivos de aprendizaje, por ejemplo, con un aprendizaje profundo. Mientras que el aprendizaje profundo puede ser verdaderamente deseable, no está implicado en un compromiso con la evaluación formativa. La evaluación formativa puede ser utilizada para ayudar a los estudiantes a alcanzar objetivos instrumentales y estrechos así como significativos y profundos.

La clarificación, comunicación y comprensión de las metas de aprendizaje y los criterios de éxito con los estudiantes tampoco establece ninguna prescripción relativa a quién determina el objetivo. Mientras los estudiantes más pequeños podrán tener relativamente poca opción acerca de qué es lo que aprenderán, cuando sean mayores asumirán más responsabilidad. Sin embargo, aún en la educación superior, donde el

estudiante elige los cursos a seguir, generalmente habrá un curriculum consensuado, de modo tal que las metas de aprendizaje actuales, y sus criterios de logro asociados, probablemente sean un asunto de negociación entre estudiantes y profesores.

Diseñar discusiones de clase efectivas, preguntas y tareas que permitan obtener evidencias acerca del aprendizaje

La segunda estrategia presentada en la Figura 1 pone el foco en la obtención de evidencias acerca de las realizaciones o logros. Si bien esta actividad tomará frecuentemente la forma de un interrogatorio, es importante señalar que puede ser incluida cualquier acción que permita obtener evidencia que pueda ser utilizada para informar la instrucción. Para los docentes de estudiantes con múltiples y profundas dificultades de aprendizaje, puede ser que la evidencia acerca del aprendizaje sea obtenida por tacto más que a través de algo reconocible como una pregunta.

El punto importante aquí es que no toda la evidencia obtenida es igualmente útil. Algunos tipos de evidencia sólo podrán apoyar funciones de monitoreo o diagnóstico. Tal como se señaló anteriormente, para que la evidencia obtenida sea instruccionalmente tratable, tanto la información obtenida como el modo en el cual ésta es obtenida, necesitan estar dirigidas por una clara comprensión de las intenciones (sean éstas definidas en forma amplia o estrecha), una comprensión del progreso en el aprendizaje (Heritage, 2008) y de las dificultades que los alumnos experimentan.
Sin embargo, sería un error asumir que las evaluaciones diagnósticas son siempre preferibles a las evaluaciones de monitoreo, y que aquellas que conducen a interpretaciones instruccionalmente tratables son siempre preferibles a las diagnósticas porque el rango de las decisiones disponibles podría ser limitado. Si la única decisión disponible es si el alumno debe repetir de grado o no, entonces será suficiente una evaluación acerca de la proporción de los aprendizajes esperados que ha sido alcanzada. Si la decisión es "¿Qué partes de esta unidad necesito revisar con la clase antes de la prueba final de la unidad?", entonces se requerirá una evaluación de mayor carácter diagnóstico.

No obstante, en general, para ser más efectiva, la instrucción necesita ser a la medida de las necesidades específicas de los alumnos individuales, y así se requerirá un mayor espectro de alternativas instruccionales que la simple repetición de las secuencias de enseñanza. Para que la evaluación formativa sea instruccionalmente tratable, el docente debe tener primero en claro la gama de movimientos instruccionales posibles y luego debe determinar qué tipo de evidencia es necesaria para tomar una decisión. En otras palabras, la elección acerca de qué tipo de evidencia recoger es guiada por una teoría del aprendizaje y casi todo el trabajo pesado intelectual está hecho antes de que el docente efectivamente pueda obtener la evidencia de las adquisiciones de los alumnos.

Proveer retroalimentación que permita a los alumnos avanzar

El requisito de retroalimentación que permita avanzar -la tercera de las

estrategias de la Figura 1- enfatiza el hecho de que la evaluación formativa efectiva es prospectiva, más que retrospectiva. Es una mirada a través del parabrisas más que a través del espejo retrovisor o, como Douglas Reeves memorablemente señaló, esta es la diferencia entre un reconocimiento médico y un post-mortem (comunicación personal, 31 de octubre de 2008). Esto sintetiza los dos hallazgos claves de Kluger y De Nisi (1996) y Hattie y Timperley (2007) discutidos antes: 1) que es más productivo pensar acerca de los procesos que son disparados por la intervención retroalimentadora, y 2) que las intervenciones retroalimentadoras tienen más probabilidades de ser efectivas si prestan atención a la tarea, al modo en que el aprendiz trabaja en la tarea y los procesos de autorregulación en los que el estudiante se implica más que si llaman la atención al *self.* Quizás, aún más simple, la retroalimentación tiene probabilidades de ser más efectiva cuando genera una respuesta de orden cognitivo más que afectivo. Por supuesto, si esto ocurre, depende no sólo de la calidad de la retroalimentación sino también del aprendiz y del contexto de aprendizaje en el cual la retroalimentación es dada y recibida (Black y Wiliam, 2005, 2009).

El otro aspecto de "la retroalimentación que hace avanzar al aprendizaje" está relacionado con los ajustes instruccionales. En lugar de dar retroalimentación al alumno, los resultados de la evaluación pueden también ofrecer retroalimentación al docente de modo tal que pueda modificar su instrucción en vistas a ser más efectivo (tanto para los estudiantes acerca de los cuales la información fue recogida o para otros estudiantes a los que se podrá enseñar en el futuro). En otras palabras, la evaluación puede ser más formativa para el docente que para el alumno.

Promover a los estudiantes como los dueños de su propio aprendizaje

Las últimas dos estrategias claves de la Figura 1 están relacionadas con el papel de los aprendices en el proceso de evaluación formativa, incluyendo el grado en que los estudiantes son dueños de su propio aprendizaje y actúan como recursos de aprendizaje para los otros, y por conveniencia están discutidos en el orden inverso en que aparecen en la Figura 1. Para que los estudiantes puedan apropiarse de su aprendizaje, ellos necesitan ser dueños de los objetivos curriculares y ser activos en la guía de su propio aprendizaje. En otras palabras, ellos deben volverse aprendices capaces de auto-regulación. La noción de aprendizaje autorregulado es un fértil foco de investigación con una vasta literatura propia, gran parte de la cual es altamente relevante para la noción de evaluación formativa. A continuación se presenta una breve síntesis de algunos de los puntos más importantes de modo tal que el lector pueda buscar información más detallada acerca de ellos.

Winne definió el aprendizaje auto-regulado como "una conducta gobernada meta- cognitivamente en la que los aprendices regulaban adaptativamente su uso de tácticas y estrategias metacognitivas en las tareas" (1996: 327). Otros han señalado que los estudiantes a menudo poseen habilidades necesarias de auto-regulación pero no las despliegan. y este problema puede ser una falta de motivación o volición (Corno, 2001).

Aun otros han argumentado acerca de la necesidad de contemplar los problemas de auto-regulación con marcos teóricos más amplios incluyendo perspectivas socioculturales (Hickey & McCaslin, 2001; McCaslin & Hickey, 2001) o socio-constructivistas (Op't Eynde, DeCorte, & Verschaffel, 2001).

Una de las definiciones más generales fue provista por Boekaerts quien define la auto-regulación como "un proceso multinivel y multicompuesto que apunta a los sentimientos, cogniciones y acciones, así como a rasgos del contexto para la modulación en el servicio de las propias metas" (2006: 347). De acuerdo con Boekaerts es difícil distinguir entre los aspectos cognitivos y motivacionales del aprendizaje auto-regulado porque está meta-cognitivamente gobernado *y* afectivamente cargado.

Se han propuesto una serie de maneras de reunir las perspectivas motivacionales y cognitivas de la autorregulación: una síntesis de ellas puede encontrarse en Wiliam (2007a). Para el propósito de este capítulo, y en términos de la estrategia particular de activar a los estudiantes como dueños de su propio aprendizaje, un modelo que es especialmente relevante es la teoría de *procesamiento dual* desarrollada por Boekaerts (1993):

> "Se asume que los estudiantes que son invitados a participar en una actividad de aprendizaje usan tres fuentes de información para formar una representación mental de la tarea en contexto y para valorarla: 1) las percepciones actuales acerca de la tarea y el contexto físico, social e instruccional en el cual ella está integrada; 2) el conocimiento de dominio específico activado y las estrategias (meta)cognitivas relativas a la tarea; y 3) las creencias motivacionales, incluyendo las capacidades de dominio específico, el interés y las creencias acerca del esfuerzo" (Boekaerts, 2006: 349).

Cuando la apreciación acerca de la tarea es positiva, la energía es activada a lo largo de la "vía del crecimiento" en la que la meta es aumentar la competencia. Boekaerts describe este tipo de auto-regulación como *top-down* porque el flujo de energía es dirigido por el estudiante. Cuando la valoración de la tarea es negativa, la atención cambia hacia la "vía del bienestar" donde la meta es evitar la amenaza, el daño o la pérdida. Esta forma de autorregulación es denominada *bottom-up* porque está provocada por pistas en el entorno más que por metas de aprendizaje. Cuando esta regulación *bottom-up* es la norma, obviamente el aprendizaje está comprometido. Sin embargo, en ciertos casos esto puede ser positivo porque al atender temporariamente al bienestar, el alumno puede encontrar un modo de cambiar su energía y atención nuevamente hacia la vía del crecimiento.

Por supuesto, la relación entre la vía *top-down* y *bottom-up* de regulación es dinámica, más que un rasgo estable de un alumno individual. Boekarts (2001) no encontró ningún vínculo directo entre las creencias motivacionales de dominio específico y la intención de aprender en ninguna de las clases de matemáticas estudiadas. Las decisiones de los estudiantes acerca de invertir esfuerzo en una actividad

de matemática dependían, en principio, de su apreciación acerca de la tarea frente a ellos; a pesar de que Ross, Rolheiser, & Hogaboam-Gray (2002) hallaron que las decisiones acerca de invertir esfuerzo estaban también influenciadas por los amigos y los padres.

Una de las mayores fortalezas del modelo de procesamiento dual es que abona la integración de una variedad de perspectivas diferentes en la idea amplia de activar a los estudiantes como dueños de su aprendizaje incluyendo la relación entre motivación e interés, el modo en que los estudiantes atribuyen sus éxitos y fracasos en el aprendizaje y la manera en que desarrollan ideas acerca de la auto-eficacia.

Por ejemplo, cuando los estudiantes están interesados en una tarea, tienen más probabilidades de comprometerse en la actividad a lo largo de la vía del crecimiento (Hidi & Harackiewicz, 2000). Cuando los estudiantes no están personalmente interesados en la tarea, el interés puede ser despertado por algún aspecto de la situación, disparando de este modo la actividad a lo largo de la vía del crecimiento. Cuando el interés no es el principal conductor de la atención, las consideraciones acerca del *valor* de la tarea versus su *costo* se volverán importantes (Eccles, Adler, Futterman, Goff, Kaczala, Meece & Midgley, 1983). En términos de las teorías de la motivación propuestas por Deci y Ryan (1994), la actividad a lo largo de la vía del crecimiento está asociada con la motivación derivada de valores dentro del individuo, mientras que la actividad a lo largo de la vía del bienestar está asociada con valores originados fuera del individuo. En términos de la teoría del logro de metas (Dweck & Leggett, 1986) los estudiantes que despliegan una orientación hacia el dominio tienen más probabilidades de activar la vía del crecimiento, mientras que los que despliegan una orientación hacia el desempeño tienen más probabilidades de activar la vía del bienestar.

Las creencias de auto-eficacia (Bandura, 1977) pueden conducir el progreso a través de cualquiera de las dos vías. A través de la vía del crecimiento la auto-eficacia conduce el empleo de estrategias cognitivas y meta-cognitivas adaptativas mientras que, a lo largo de la vía del bienestar, es probable que las creencias de auto-eficacia alejen a los estudiantes de las metas de eludir el desempeño y hacia metas enfocadas a lograr el desempeño. Visiones similares de la habilidad como *incremental* (Dweck, 2000) ayudan al estudiante a mantenerse en la vía del crecimiento, mientras las visiones de la habilidad como *entidad* dirigen la actividad hacia la vía del bienestar donde los detalles de la actividad-en-el-contexto, apreciadas a la luz de las visiones acerca de la propia capacidad, influenciarán las decisiones acerca de comprometerse o no en la tarea.

Activar a los estudiantes como recursos unos para los otros

La estrategia final presentada en la Figura 1 es la activación de los estudiantes como posibles recursos unos para los otros. De alguna manera esta estrategia provee un foco para las otras cuatro estrategias, en la medida en que combina aspectos de

cada una de ellas. Para que los estudiantes puedan evaluar el trabajo de los otros, deben haber internalizado las metas pedagógicas o los criterios de realización y estas comprensiones deben estar disponibles para los estudiantes para que puedan ser utilizadas en sus propias producciones (Black *et al*, 2003). Además, dado que evaluar el trabajo de otro implica menor carga emocional que el intento de evaluar el propio trabajo, la evaluación de los pares provee una plataforma útil para una auto evaluación efectiva y por lo tanto para mejorar la auto-regulación en el aprendizaje (Black *et al*, 2003: 62). En la tutoría de pares y en otras formas de aprendizaje colaborativo, el par frecuentemente desempeña el papel del profesor, de modo que la obtención de la evidencia y la provisión de retroalimentación son fundamentales. Verdaderamente, los límites entre las estrategias frecuentemente se borran. Cuando los docentes les piden a los estudiantes que revisen su aprendizaje construyendo ítems de prueba (con sus correspondientes respuestas correctas) como en el estudio de Foos, Mora y Tkacz (1994), los estudiantes necesitan pensar cuidadosamente acerca de los objetivos de aprendizaje del trabajo que han estado estudiando, y acerca de cuál es una buena manera de obtener evidencia. Cuando estos ítems son administrados a otros estudiantes (Fontana & Fernandes, 1994) los alumnos se convierten en recursos de aprendizajes los unos para los otros, y por consiguiente, mejoran sus propias habilidades de auto-regulación.

Resumen

Este capítulo ofreció una breve historia de la idea de evaluación formativa con una reseña de la investigación que apoya su eficacia en contextos educativos. Aunque existen problemas metodológicos inevitables para sintetizar los resultados provenientes de estudios que usan diferentes instrumentos para medir resultados y son llevados a cabo en el marco de diferentes tradiciones, no hay dudas de que el aumento del uso de la evaluación formativa es una de las maneras más efectivas y más rentables de mejorar el desempeño de los estudiantes. Más aún, el efecto parece ser generalizable a través de diferentes tipos de aprendizaje, en un amplio rango de contextos y en alumnos de todas las edades.

A medida que la idea de evaluación formativa se ha desarrollado, la definición del término ”formativa” ha fluctuado desde una descripción del momento de una evaluación (cualquier evaluación previa a “la gran evaluación”) hacia una descripción de un tipo de instrumento. Sin embargo, dado que la evidencia proveniente de un instrumento de evaluación puede ser usada en una serie de maneras posibles, este capítulo ha propuesto una definición de evaluación formativa en términos del grado en que la evidencia acerca de los logros de los estudiantes es usada para informar las decisiones acerca de la enseñanza y el aprendizaje. En particular, la evaluación formativa se refiere a la creación y capitalización de momentos de contingencia en la instrucción (incluyendo tanto la enseñanza como el aprendizaje) en vistas a regular el proceso de aprendizaje de modo más efectivo.

Si bien resulta algo abstracta su formulación, esta definición abona por su inmediata aplicación en contextos educativos en términos de cinco estrategias claves:

1. clarificar, compartir y comprender las metas de aprendizaje y los criterios de logro;
2. diseñar discusiones de clase efectivas, preguntas y tareas que permitan obtener evidencias acerca del aprendizaje;
3. proveer retroalimentación que permita a los alumnos avanzar;
4. promover a los estudiantes como recursos de enseñanza para otros alumnos;
5. promover a los estudiantes como los dueños de su propio aprendizaje.

Las cinco estrategias no son sólo importantes procesos en la instrucción sino que parecen ser dos poderosas lentes para pensar acerca de las prácticas (cita proveniente del trabajo con docentes) y, por lo tanto, para ayudar a los docentes a involucrarse con las problemáticas más amplias de la psicología, el curriculum y la pedagogía.

Recomendaciones para el trabajo futuro

Tal como Kluger y De Nisi (1996) sugirieron, probablemente no sean de utilidad próximos diseños para identificar con mayor precisión el tamaño del impacto en el aprendizaje que puede ser alcanzado mediante la evaluación formativa. Lo que *sí sería* probablemente provechoso es el desarrollo de estudios que vincularan las distintas maneras de intervenciones retroalimentadoras con los procesos de aprendizaje que ellas generan. Esos estudios, llevados a cabo a lo largo de largos períodos de tiempo (por lo menos un año), podrían también mostrar si la instrucción de buena calidad es compatible con un incremento en el éxito en las pruebas estandarizadas, lo que sería importante para comprender cómo mejorar la enseñanza en contextos que hacen un amplio uso de las pruebas que son utilizadas para conservar la responsabilidad de estudiantes y docentes. Sin esa evidencia, los intentos de reforma probablemente se encuentren con la siguiente reacción: "Me encantaría enseñar para una comprensión más profunda, pero tengo que mejorar los puntajes en las pruebas".

Sin embargo, es posible que estos estudios sean en última instancia mucho menos importante que los estudios acerca de cómo apoyar a los docentes para hacer un mayor uso de la evaluación formativa en su propia práctica. Si bien no sabemos nada acerca de qué es lo que define a los usos más efectivos de la evaluación formativa, probablemente sabemos suficiente para construir un consenso sustancial alrededor de qué tipo de clases queremos. Sabemos mucho menos acerca de cómo tener más de esas clases. Como señalaron Black y Wiliam:

> "Es difícil ver cómo cualquier innovación en materia de evaluación formativa puede ser tratada como un cambio marginal en el trabajo de la clase. Todo ese trabajo involucra algún grado de retroalimentación entre aquellos que son enseñados y el docente, y esto está implicado en la calidad de sus interacciones, las cuales son el corazón de la pedagogía" (1998a: 16).

Hay muchas historias exitosas (por ejemplo Wiliam, Lee, Harrison & Black, 2004), pero sabemos muy poco acerca de los factores que sostienen la implementación de innovaciones educativas a escala (Coburn, 2003; Thompson & Wiliam, 2008). Si queremos asegurar las mejoras en los resultados educativos que la investigación existente sobre la evaluación formativa ha demostrado como posible, el diseño de formas de apoyo a los profesores para desarrollar sus prácticas de evaluación formativa a escala, debe ser la prioridad principal.

* N del T: La expresión en inglés es EXIT PASS. Se trata de una herramienta rápida de evaluación en la que los alumnos responden a una pregunta sustantiva planteada por el docente al final de la clase. Lo hemos traducido como permiso de salida, pero esta expresión no se utiliza en español.

Notas

[*] WILIAM, D. (2009) "An Integrative Summary of the Research Literature and Implications for a New Theory of Formative Assessment", en ANDRADE, H. and CIZEK, G. J. (eds.) *Handbook of Formative Assessment,* New York and London, Routledge

Handbook of Formative Assessment by Dylan Wiliam . Copyright 2009 by TAYLOR & FRANCIS GROUP LLC - BOOKS. Reproduced with permission of TAYLOR & FRANCIS GROUP LLC - BOOKS in the format Journal via Copyright Clearance Center

Bibliografía

ALEXANDER, R., *Essays on pedagogy*. York, UK: Diálogos, 2008.

ALLAL, L., & LÓPEZ, L. M., Formative assessment of learning: a review of publications in French. In J. Looney (Ed.), *Formative assessment: improving learning in secondary classrooms.* Paris, France: Organisation for Economic Cooperation and Development, 2005.

BANDURA, A., Self-efficacy: towards a unifying theory of behavioral change. *Psychological Review, 84*(2), 1997.

BANGERT-DROWNS, R. L., KULIK, C.-L. C., KULIK, J. A., & MORGAN, M. T., The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 1991.

BLACK, P. J. & WILIAM, D., Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 1998a.

BLACK, P. J., & WILIAM, D, Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 1998b.

BLACK, P., & WILIAM, D.,Developing a theory of formative assessment. In J. Gardner (Ed.), *Assessment and learning*. London, UK: Sage, 2005.

BLACK, P. J., & WILIAM, D. (in press). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability, 1*(1).

BLACK, P., HARRISON, C., LEE, C., MARSHALL, B., & WILIAM, D., *Assessment for learning: Putting it into practice*. Buckingham, UK: Open University Press, 2003.

BLACK, P., HARRISON, C., LEE, C., MARSHALL, B., & WILIAM, D., Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 2004.

BLOOM, B. S., The search for methods of instruction as effective as one-to-one tutoring. *Educational Leadership, 41*(8), 1984.

Boekaerts, M., Being concerned with well being and with learning. *Educational Psychologist, 28*(2), 1993.

BOEKAERTS, M., Context sensitivity: Activated motivational beliefs, current concerns and emotional arousal. In S. Volet & S. Järvelä (Eds.), *Motivation in learning contexts: Theoretical advances and methodological implications*. Oxford, UK: Pergamon, 2001.

BOEKAERTS, M., Self-regulation and effort investment. In K. A. Renninger & I. E. Sigel (Eds.), *Handbook of child psychology volume 4: Child psychology in practice,* (6 ed.) New York: Wiley, 2006.

BROADFOOT, P. M., DAUGHERTY, R., GARDNER, J., GIPPS, C. V., HARLEN, W., JAMES, M., et al., *Assessment for learning: Beyond the black box*. Cambridge, UK: University of Cambridge School of Education, 1999.

BROADFOOT, P. M., DAUGHERTY, R., GARDNER, J., HARLEN, W., JAMES, M., & STOBART, G., *Assessment for learning: 10 principles*. Cambridge, UK: University of Cambridge School of Education, 2002.

BROOKHART, S. M., Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record, 106*(3), 2004.

BROOKHART, S. M., Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative classroom assessment: Theory into practice*. New York: Teachers College Press, 2007.

COBURN, C., Rethinking scale: moving beyond numbers to deep and lasting change. *Educational Researcher,* 32(6), 2003.

CORNO, L., Volitional aspects of self-regulated learning. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated leaning and academic achievement: Theoretical perspectives,* (2 ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, 2001.

COWIE, B., & BELL, B., A model of formative assessment in science education. *Assessment in Education: Principles, Policy, and Practice, 6*(1), 1999.

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

DECI, E. L., & RYAN, R. M., Promoting self-determined education. *Scandinavian Journal of Educational Research,* 38(1), 1994.

DEMPSTER, F. N., Synthesis of research on reviews and tests. *Educational Leadership, 48*(7), 1991.

DEMPSTER, F. N., Using tests to promote learning: A neglected classroom resource. *Journal of Research and Development in Education, 25*(4), 1992.

DENVIR, B., & BROWN, M. L., Understanding of number concepts in low-attaining

7-9 year olds: Part 1. Development of descriptive framework and diagnostic instrument. *Educational Studies in Mathematics,* 17(1), 1986a.

DENVIR, B., & BROWN, M. L., Understanding of number concepts in low-attaining 7-9 year olds: Part II. The teaching studies. *Educational Studies in Mathematics,* 17(2), 1986b.

DWECK, C. S., *Self-theories: Their role in motivation, personality and development*. Philadelphia: Psychology Press, 2000.

DWECK, C. S., & LEGGETT, E. L., Motivational processes affecting learning. *American Psychologist (Special Issue: Psychological Science and Education), 41(*10), 1986.

ECCLES, J. S., ADLER, T. F., FUTTERMAN, R., GOFF, S. B., KACZALA, C. M., MEECE, J. L., et al., Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motivation*. San Francisco: W H Freeman, 1983.

ELSHOUT-MOHR, M., Feedback in self-instruction. *European Education, 26*(2), 1994.

FONTANA, D., & FERNANDES, M., Improvements in mathematics performance as a consequence of self-assessment in Portugese primary school pupils. *British Journal of Educational Psychology, 64*(4), 1994.

FOOS, P. W., MORA, J., & TKACZ, S., Student study techniques and the generation effect. *Journal of Educational Psychology, 86*(4), 1994.

FORSTER, M., & MASTERS, G. N., Bridging the conceptual gap between classroom assessment and accountability. In M. Wilson (Ed.), *Towards coherence between classroom assessment and system accountability: 103rd Yearbook of the National Society for the Study of Education* (Part II). Chicago, IL: University of Chicago Press, 2004.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation - a meta-analysis. *Exceptional children, 53*(3), 199-208.

GIPPS, C. V., & STOBART, G., *Assessment: A teacher's guide to the issues* (3rd ed.). London, UK: Hodder and Stoughton, 1997.

HATTIE, J., & TIMPERLEY, H., The power of feedback. *Review of Educational Research, 77*(1), 2007.

HERITAGE, M., *Learning progressions: supporting instruction and formative assessment*. Washington, DC: Council of Chief State School Officers, 2008.

HICKEY, D. T., & McCASLIN, M., A comparative, sociocultural analysis of context and motivation. In S. Volet & S. Järvelä (Eds.), *Motivation in learning contexts*. Oxford, UK: Pergamon, 2001.

HIDI, S., & HARACKIEWICZ, J. M., Motivating the academically umotivated: A critical issue for the 21st century. *Review of Educational Research, 70*(2), 2000.

JAMES, M., *Assessment for learning.* Annual Conference of the Association for Supervision and Curriculum Development (Assembly session on 'Critique of Reforms in Assessment and Testing in Britain') held in New Orleans, LA. Cambridge, UK: University of Cambridge Institute of Education, 1992.

KAHL, S., Where in the world are formative tests? Right under your nose! *Education Week,* 25(4), 2005.

KLUGER, A. N. & DeNISI, A., The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 1996.

KÖLLER, O., Formative assessment in classrooms: A review of the empirical German literature. In J. Looney (Ed.), *Formative assessment: Improving learning in secondary classrooms*. Paris, France: Organisation for Economic Cooperation and Development, 2005.

LEWIS, C. C., *Lesson study: A handbook of teacher-led instructional change*. Philadelphia: Research for Better Schools, 2002.

LOONEY, J. (Ed.)., *Formative assessment: Improving learning in secondary classrooms*. Paris, France: Organisation for Economic Cooperation and Development, 2005.

McCASLIN, M., & HICKEY, D. T. (2001). Educational psychology, social constructivism, and educational practice: A case of emergent identity. *Educational Psychologist, 36*(2), 133-140.

MITCHELL, R., *Testing for learning*. New York USA: Free Press Macmillan, 1992.

National Assessment of Educational Progress. *The nation's report card: Mathematics 2005* (Vol. NCES). Washington, DC: Institute of Education Sciences, 2006.

NATRIELLO, G., The impact of evaluation processes on students. *Educational Psychologist,* 22(2), 1987.

NYQUIST, J. B., *The benefits of reconstruing feedback as a larger system of formative assessment: A meta-analysis.* Unpublished master's thesis. Nashville, TN: Vanderbilt University, 2003.

OP'T EYNDE, P., DECORTE, E., & VERSCHAFFEL, L., "What to learn from what we feel?" The role of students' emotions in the mathematics classroom. In S. Volet & S. Järvelä (Eds.), *Motivation in learning contexts: Theoretical advances and methodological implications*. Oxford, UK: Pergamon, 2001.

POPHAM, W. J., Phony formative assessments: Buyer beware! *Educational Leadership, 64*(3), 2006.

POPHAM, W. J., *Determining the instructional sensitivity of accountability tests.* Paper presented at a Symposium entitled "Three practical policy-focused procedures for determining an accountability test's instructional sensitivity" at the annual conference of the American Educational Research Association, Chicago, IL, 2007.

PRYOR, J., & CROSSOUARD, B., *A sociocultural theorization of formative assessment.* Paper presented at Sociocultural Theory in Educational Research and Practice Conference held at University of Manchester.Brighton, UK: University of Sussex, 2005.

RAMAPRASAD, A., On the definition of feedback. *Behavioural Science, 28*(1), 1983.

ROSS, J. A., ROLHEISER, C., & HOGABOAM-GRAY, A., Influences on student

cognitions about evaluation. *Assessment in Education: Principles, Policy, and Practice, 9*(1), 2002.

RUIZ-PRIMO, M. A., SHAVELSON, R. J., HAMILTON, L., & KLEIN, S., On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching, 39*(5), 2002.

SADLER, D. R., Formative assessment and the design of instructional systems. *Instructional Science, 18,* 1989.

SHEPARD, L. A., Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates, 2007.

SHEPARD, L. A., HAMMERNESS, K., DARLING-HAMMOND, L., RUST, F., SNOWDEN, J. B., GORDON, et al., Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do*. San Francisco, CA: Jossey-Bass, 2005.

SHUTE, V. J., Focus on formative feedback. *Review of Educational Research,* 78(1), 2008.

STIGGINS, R. J., Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 2002.

SUTTON, R., *Assessment for learning*. Salford, UK: RS Publications, 1995.

THOMPSON, M., & WILIAM, D., Tight but loose: A conceptual framework for scaling up school reforms. In E. C. Wylie (Ed.), *Tight but loose: Scaling up teacher professional development in diverse contexts* (RR-08-29). Princeton, NJ: Educational Testing Service, 2008.

VINNER, S., From intuition to inhibition—mathematics, education and other endangered species. In E. Pehkonen (Ed.), *Proceedings of the 21st conference of the International Group for the Psychology of Mathematics Education* (Vol. 1). Lahti, Finland: University of Helsinki Lahti Research and Training Center, 1997.

WIENER, N., *Cybernetics, or the control and communication in the animal and the machine*. New York: John Wiley, 1948.

WILIAM, D., LEE, C., HARRISON, C., & BLACK, P. J., Teachers developing assessment for learning: impact on student achievement. *Assessment in Education: Principles Policy and Practice,* 11(1), 2004.

WILIAM, D., Keeping learning on track: Classroom assessment and the regulation of learning. In F. K. Lester Jr (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053-1098). Greenwich, CT: Information Age Publishing, 2007a.

WILIAM, D., *An index of sensitivity to instruction* Paper presented at a Symposium entitled "Three practical policy-focused procedures for determining an accountability test's instructional sensitivity" at the annual conference of the American Educational Research Association held at Chicago, IL, 2007b.

WILIAM, D., Content *then* process: teacher learning communities in the service of formative assessment. In D. B. Reeves (Ed.), *Ahead of the curve: The power*

of assessment to transform teaching and learning. Bloomington, IN: Solution Tree, 2007c.

WILIAM, D., International comparisons and sensitivity to instruction. *Assessment in Education: Principles, Policy, and Practice,* 15(3), 2008.

WILIAM, D., & BLACK, P. J., Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal,* 22(5), 1996.

WILIAM, D., & THOMPSON, M., Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning*. Mahwah, NJ: Lawrence Erlbaum Associates, 2007.

WINNE, P. H., A metacognitive view of individual differences in self-regulated learning. *Learning and Individual Differences, 8*, 1996.