Para citar este artículo, le recomendamos el siguiente formato:

Martínez Rizo, F. (2012). Investigación empírica sobre el impacto de la evaluación formativa. Revisión de literatura. *Revista Electrónica de Investigación Educativa, 14*(1), 1-15. Consultado en http://redie.uabc.mx/vol14no1/contenido-martinezrizo12.html

ISSN: 1607-4041

Revista Electrónica de Investigación Educativa

Volumen 14, Núm. 1, 2012

Investigación empírica sobre el impacto de la evaluación formativa: Revisión de literatura

An Empirical Study on the Impact of Formative Assessment: A Literature Review

Felipe Martínez Rizo
fmrizo@prodigy.net.mx
Universidad Autónoma de Aguascalientes

San Cosme 108, C.P. 20010
Aguascalientes, Aguascalientes, México

(Recibido: 4 de noviembre de 2011; aceptado para su publicación: 2 de febrero de 2012)

Resumen

En muchos sistemas educativos hay un creciente interés por el enfoque de la evaluación del aprendizaje que designan expresiones como evaluación formativa, evaluación en aula o evaluación para el aprendizaje. Este interés suele basarse en opiniones muy positivas sobre dicho enfoque difundidas primero por los partidarios del sistema Mastery Learning desarrollado por Benjamin S. Bloom y, desde fines de la década de 1990, por otros estudiosos. El sustento empírico de esas posturas, sin embargo, no siempre es sólido, lo que tampoco debe entenderse como una descalificación del enfoque. En este artículo se presenta una revisión de la literatura al respecto, en la que puede basarse una opinión mejor informada.

Palabras clave: Evaluación del aprendizaje, educación básica, estado del conocimiento.

Abstract

In many educational systems there is a growing interest in the approach to learning assessment referred to as formative assessment, classroom assessment or assessment for learning. This interest is usually based on the very positive opinions spread initially by supporters of the Mastery Learning model developed by Benjamin S. Bloom and, since the 1990s, by other scholars. Empirical support for these views, however, is not always very strong, which does not necessarily mean that the approach is wrong. This article presents a review of the literature on the subject, providing a foundation for a better informed position.

Key words: Performance based assessment, elementary education, state of the art.

I. Introducción

La noción de evaluación formativa (EF) denota las acciones hechas para valorar el avance de los alumnos en el desarrollo de conocimientos o competencias, buscando aportar elementos al maestro y a los alumnos mismos para adecuar sus esfuerzos para alcanzar las metas de aprendizaje establecidas, y no para emitir un juicio definitivo al respecto. La noción opuesta es la de evaluación sumativa.

Aunque reciente, la literatura sobre EF comienza a proliferar y, en general, parte del supuesto de que su uso produce beneficios importantes sobre el aprendizaje. Sin embargo los estudios de carácter empírico que sometan este supuesto a la prueba de la experiencia son escasos. En estas páginas se revisa la literatura al respecto, con énfasis en la producida en los medios anglosajones.

Algunas síntesis de estudios sobre el sistema Mastery Learning que promovió Benjamín S. Bloom en las décadas de 1960 y 1970, llegaban a la conclusión de que era "una de las estrategias de enseñanza más efectivas que los maestros pueden utilizar, en cualquier nivel educativo" (Walberg, 1984, en Guskey 2007, p. 75).

Bloom plantea el tema en términos de lo que llama el problema de las dos sigmas, en alusión al dato de que la diferencia entre los alumnos de alto y bajo rendimiento suele situarse en el rango de dos desviaciones estándar; esto quiere decir que el reto de reducir las brechas del rendimiento de manera significativa en un sistema educativo consiste en mejorar el desempeño de los alumnos de menor rendimiento en el equivalente a dos desviaciones estándar (dos sigmas).

Según Bloom es posible conseguir mejoras de esa importancia con sistemas de enseñanza que asignan un tutor a cada alumno de bajo rendimiento (enseñanza tutorial uno a uno), lo cual es costoso; pero él afirma que con su sistema es posible obtener resultados similares, con un costo que no difiere mucho del que implican las formas tradicionales de enseñanza (Bloom, 1984a y 1984b).

II. Estudios en perspectiva optimista

Sin contar las relativas a resultados del modelo Mastery Learning, seguramente la revisión de literatura que más ha influido en las ideas sobre el efecto de la EF es la que publicaron en 1998 Paul Black y Dylan Wiliam, en el número 1 de 1998 de la revista Assessment in Education: principles, policy & practice. El artículo no es el primero en ese sentido, puesto que revisa trabajos publicados antes, entre 1988 y 1997; sin embargo, la conclusión tan positiva a la que llega llamó la atención entre los interesados en el tema, algunos de los cuales lo retomaron sin considerar sus alcances y límites, lo que posiblemente dio lugar a que se extendiera una visión acrítica de la evaluación formativa, en un medio que a veces parece ávido de soluciones milagrosas para los problemas que enfrenta.

Black y Wiliam tomaron como línea de base las revisiones de Natriello (1987) y Crooks (1988) y utilizaron otras revisiones (Black, 1993; Bangert-Drowns, Kulik, Kulik y Morgan, 1991a y 1991b; Kulik, Kulik y Bangert-Drowns, 1990); consultaron el ERIC *(*Education Resources Information Center), rastrearon referencias citadas en las ya localizadas (snowball approach); e hicieron una búsqueda de artículos en 76 revistas. De esta manera identificaron 681 publicaciones que parecían relevantes a primera vista, que luego se redujeron a unas 250.

La revisión se organizó en siete secciones: ejemplos notables; evaluación por los maestros; perspectiva de los alumnos; papel del maestro; estrategias y tácticas usados por los maestros; sistemas en que la evaluación formativa tiene un papel especial; y retroalimentación. En cada sección los textos identificados se revisan con diferente amplitud, pero la primera, la de ejemplos notables, atrae la atención tanto por su ubicación al inicio del artículo como por la naturaleza positiva de las conclusiones a que llegan los autores, que sustentan las afirmaciones reiteradas y enfáticas que se hacen. Así, en el resumen del trabajo, Black y Wiliam dicen:

> Varios estudios muestran evidencia firme de que las innovaciones que se diseñan para reforzar la retroalimentación frecuente que el alumno recibe sobre su aprendizaje produce ganancias substanciales (…) (1998, p. 7)

La idea se retoma con fuerza similar al final del trabajo, donde los autores responden la pregunta sobre las implicaciones para las políticas de sus hallazgos:

> La investigación reportada muestra en forma concluyente que la EF mejora el aprendizaje. Las ganancias en desempeño parecen muy considerables y son de las más grandes reportadas para una intervención educativa. Como ejemplo de su importancia, si se alcanzara a escala nacional un efecto de 0.7 (*size effect*), equivaldría a elevar el puntaje promedio en matemáticas de un país promedio como Inglaterra, Nueva Zelanda o Estados Unidos, al nivel de los cinco mejores, detrás de los países de la cuenca del Pacífico como Singapur, Corea, Japón y Hong Kong (…) (Black y Wiliam, 1998, p. 61)

Los estudios seleccionados por Black y Wiliam para incluir en la primera sección de su trabajo, de ejemplos destacados, son ocho: 1) Un proyecto que involucró a

25 profesores portugueses de matemáticas, con 246 alumnos de ocho y nueve años de edad y 108 más de 10 a 14 años. 2) La experiencia de un profesor a lo largo de 18 años durante los cuales utilizó el modelo de *Mastery Learning* en sus cursos, con unos 7,000 estudiantes. 3) Otro estudio que utilizó el modelo de Mastery Learning, con 120 estudiantes universitarios estadounidenses, en cuatro grupos en un diseño 2 x 2. 4) Uno más con 838 niños de cinco años de edad de medio desfavorecido, distribuidos en un grupo experimental y uno de control. 5) Un experimento con 48 alumnos de 11 años de edad, de 12 grupos en cuatro escuelas de Israel, seleccionados de manera que la mitad fueran del cuartil superior y el resto del inferior, en matemáticas y lengua. 6) Un estudio con 44 alumnos de 9 o 10 años de edad en una escuela elemental de los Estados Unidos. 7) Un trabajo con 12 grupos de 30 alumnos cada uno, en dos escuelas estadounidenses de educación media. 8) Un meta-análisis de 21 estudios con alumnos de preescolar a enseñanza media superior con necesidades educativas especiales de importancia media.

Black y Wiliam advierten sobre las limitaciones de los estudios revisados. Como ejemplo, en relación con una revisión sobre la efectividad de la retroalimentación (Kluger y De Nisi, 1996) señalan que, de más de 3,000 reportes analizados, la gran mayoría debieron descartarse por fallas metodológicas como falta de controles adecuados, mezcla de efectos de la retroalimentación con otros, número reducido de sujetos (<10), ausencia de mediciones del rendimiento y datos insuficientes para estimar el tamaño del efecto. Solamente se conservaron 131 reportes que no presentaban las fallas anteriores (Black y Wiliam, 1998, p. 48)

Se mencionan efectos de la retroalimentación en sentidos opuestos, según se refiera a la tarea o a la persona. Un trabajo reporta que la retroalimentación que se refiere a la persona parece tener efectos negativos sobre el desempeño y otro que los maestros eficaces elogian menos a sus alumnos que el docente promedio, lo que coincide con otros hallazgos de que los elogios verbales y la retroalimentación de apoyo a la persona puede aumentar el interés y mejorar actitudes del alumno, pero tiene poco o nulo impacto sobre el desempeño (Black y Wiliam, 1998, p. 49-50).

La descripción que hacen Black y Wiliam de los ocho ejemplos notables suscita dudas en cuanto a la solidez de conclusiones tan contundentes como las citadas, pues parece difícil llegar a ellas sin muchas salvedades, a partir de una gama bastante reducida de trabajos diferentes, algunos de los cuales presentan claras debilidades. Pese a ello, la heterogeneidad misma de los ejemplos es manejada por los autores citados como argumento a favor de su punto de vista:

> (…) Pese a la existencia de algunos resultados marginales e incluso negativos, el rango de las condiciones y contextos en los que los estudios revisados han mostrado que se pueden alcanzar tales ganancias debe indicar que los principios que subyacen al logro de mejoras sustanciales en el aprendizaje son robustos (…) (Black y Wiliam, 1998, p. 61)

Al final de su revisión, Black y Wiliam (1998) advierten a los lectores sobre la dificultad que supone modificar en profundidad prácticas muy arraigadas:

> (…) de esta revisión no emerge un modelo óptimo en que se pueda sustentar una política. Lo que emerge son principios orientadores, con la advertencia de que los cambios requeridos en la práctica docente son centrales y no marginales, y deben ser incorporados por cada docente a su propia práctica en la manera propia de cada uno. En otras palabras, una reforma de tales dimensiones inevitablemente llevará mucho tiempo y requerirá el continuo apoyo de educadores e investigadores. (p. 62)

En forma muy clara, el trabajo citado señala:

> Sería deseable, y se podría esperar como lo habitual, que una revisión como ésta tratara de hacer un meta-análisis de los estudios cuantitativos revisados. El que esto difícilmente parezca posible lleva a reflexionar sobre este campo de investigación. Esta revisión aprovechó material útil de varios estudios basados en meta-análisis; éstos, sin embargo, centraban la atención en aspectos bastante restringidos de la evaluación formativa, por ejemplo la frecuencia con la que se formulan preguntas. El valor de sus generalizaciones es también dudoso porque se ignoran aspectos clave de los estudios sintetizados, por ejemplo la calidad de las preguntas que se formulan, ya que la mayoría de los investigadores no ofrecen evidencias sobre estos puntos.
>
> Hay estudios cuantitativos que exploran la evaluación formativa de manera más comprensiva, y algunos se discuten en el texto, pero el número con un rigor cuantitativo adecuado y comparable debe situarse, como máximo, en el orden de 20. Sin embargo, si bien cada estudio es riguroso dentro de su propio marco y en relación son sus objetivos, y aunque muestran cierta coherencia en lo que se refiere a las ganancias de aprendizaje asociadas con las iniciativas de evaluación en aula, las diferencias subyacentes entre los estudios son tales que cualquier agregación de sus resultados tendría poco sentido. (Black y Wiliam, 1998, p. 52-53)

Pese a lo anterior, varias lecturas del texto al que se refieren estos comentarios han retomado únicamente las conclusiones favorables, sin matiz alguno, e incluso contradiciendo afirmaciones expresas, como en el caso siguiente:

> Con base en su síntesis de más de 250 artículos [Black y Wiliam] reportan que la respuesta [a la pregunta sobre si hay evidencias de que mejorar la calidad de la EF eleva el rendimiento de los alumnos] es un rotundo sí. De esas fuentes, unas 40 responden la pregunta con diseños experimentales suficientemente rigurosos para permitir la agregación de los datos para hacer un meta-análisis que permita estimar el efecto atribuible a EF mejoradas sobre el puntaje en pruebas sumativas. (Stiggins, 2001, p. 10)

La diferencia entre lo que dice el texto de Black y Wiliam y la lectura de Stiggins es notable y hace parecer excesivo el tono optimista de esa y otras interpretaciones.

Un importante trabajo sobre evaluación formativa publicado por la Organización para la Cooperación y el Desarrollo Económicos (OCDE) retoma el texto ya citado de las conclusiones del artículo de Black y Wiliam:

> (…) la evaluación formativa mejora el aprendizaje. Las ganancias en el desempeño parecen muy considerables y, como se ha señalado, son de las más grandes reportadas para una intervención educativa. (Centre for Educational Research and Innovation, 2005, p. 22)

El trabajo de la OCDE, sin embargo, matiza la afirmación anterior como sigue:

> Si bien la EF no es una solución mágica (silver bullet) que puede resolver todos los retos educativos, es un medio poderoso para alcanzar el objetivo de resultados de alto desempeño y alta equidad, y ofrece a los alumnos el conocimiento y las habilidades para seguir aprendiendo a lo largo de la vida. Los sistemas educativos que enfrenten las tensiones que impiden una práctica más amplia de la EF y fomenten culturas de evaluación probablemente avanzarán mucho más hacia tales metas. (CERI, 2005, p. 27)

En la Conferencia Internacional sobre Evaluación para el Aprendizaje que tuvo lugar en Chester en 2001, se llegó a la conclusión de que las discusiones sobre las prioridades de investigación en torno al tema:

> (…) se desarrollaron con plena conciencia del hecho de que contamos ya con evidencia convincente, basada en investigaciones, en cuanto al impacto de la "enseñanza para el aprendizaje" sobre el rendimiento de los alumnos: se pueden conseguir avances sin precedentes. Tenemos también evidencia convincente, basada en investigaciones, sobre la baja calidad de muchas evaluaciones que se hacen en el aula, debido a la persistente falta de oportunidades que tienen los maestros para desarrollar sus competencias de evaluación (*assessment literacy*). (Stiggins y Arter, 2002, p. 3)

En un texto más reciente, Stiggins (2007) sigue mostrando su perspectiva optimista: "La evidencia recolectada en todo el mundo revela de manera consistente efectos directamente atribuibles a la aplicación efectiva de EF en aula, que van de media a una y media desviación estándar" (p. 18).

Stiggins dice que Bloom (1984a) reportaba avances de una a dos desviaciones estándar gracias a la aplicación de su modelo de mastery learning; menciona las ganancias de 0.5 a una desviación estándar reportadas por Black y Wiliam según la revisión mencionada; cita el trabajo de Meisels, Atkins-Burnett, Xue, DiPrima y Son (2003), con ganancias de 1 a 1.5 desviaciones; y retoma el trabajo de Rodríguez (2004), con base en los resultados de la aplicación del Estudio Internacional de Tendencias en Matemáticas y Ciencias (TIMSS, por sus siglas en inglés) en los Estados Unidos. Para terminar, dice:

> Según estos investigadores, los avances esperados en las puntuaciones de desempeño rivalizan con la implementación de sistemas de enseñanza tutorial uno a uno en cuanto a su impacto en el rendimiento de los alumnos, además de que

> las mayores ganancias son conseguidas por los de menor desempeño, con lo que las brechas se reducen. (Stiggins, 2007, p. 19)

Otras revisiones de las que se extraen conclusiones favorables para la evaluación formativa se refieren a los efectos de la retroalimentación. Marzano presenta así algunos trabajos sobre el tema:

> Como resultado de revisar casi 8,000 estudios, Hattie (1992) encontró que, sin duda, "la modificación singular más poderosa para mejorar rendimiento es la retroalimentación". La receta más simple para mejorar la educación es "cucharadas de retroalimentación". Más recientemente, Hattie y Timperley (2007) actualizaron y ampliaron la revisión sobre retroalimentación y llegaron a la misma conclusión. Desafortunadamente no todas las formas de retroalimentación son igualmente efectivas. Un meta-análisis de Bangert-Drowns, Kulik, Kulik y Morgan (1991) que revisó los hallazgos de 40 estudios sobre evaluación en aula, encontró que decir simplemente al alumno si sus respuestas son correctas o incorrectas tenía efecto negativo sobre el aprendizaje, mientras que explicar la respuesta correcta y/o pedir que siguiera mejorando sus respuestas se asociaba con ganancias de 20 puntos percentilares en el desempeño. (Marzano, 2007, p. 103-104)

III. Perspectivas críticas

La experiencia de la complejidad de los fenómenos educativos y la dificultad de introducir cambios que produzcan consecuencias importantes hace tomar con reservas los textos de la sección anterior, que a veces parecen promover una panacea más que, tras cierto tiempo, provocará una desilusión tanto más fuerte cuanto mayores hubieran sido las expectativas inicialmente despertadas.

Esta idea se ve reforzada por las salvedades que contienen los mismos textos citados, que una lectura atenta no deja de advertir, y que en ocasiones se incluyen de manera tan expresa que sorprende que no sean atendidas por algunas lecturas posteriores. Confirma y refuerza la reflexión crítica sobre conclusiones demasiado optimistas un trabajo reciente de dos estudiosos de la Universidad de Arkansas, publicado con el título Una revisión crítica de la investigación sobre evaluación formativa. La limitada evidencia científica del impacto de la evaluación formativa en la educación. Al principio de su texto los autores señalan que:

> Una creencia casi nunca cuestionada es que la investigación demuestra en forma concluyente que el uso de evaluación formativa facilita la mejora de las prácticas de enseñanza, identifica lagunas en el currículo y contribuye a aumentar el desempeño de los alumnos. Sin embargo… una revisión de la literatura reveló la limitada evidencia empírica que demuestra que el uso de evaluación formativa en el aula resulta directamente en cambios marcados en los resultados educativos. (Dunn y Mulvenon, 2009, p. 1)

El texto comienza con una discusión sobre la forma en que se suele definir la noción de EF, los autores indica que la heterogeneidad al respecto es muy considerable, de manera que la tarea de analizar en forma rigurosa su posible

impacto se dificulta mucho. El artículo analiza en particular la revisión de literatura hecha por Black y Wiliam, a la que se refiere el apartado anterior de este artículo, cuya influencia se puede apreciar por el elevado número de veces que se le cita en la revistas académicas (194 según el Social Science Index revisado por Dunn y Mulvenon (2009, p. 5).

La revisión de los ocho estudios que Black y Wiliam utilizan para sustentar sus conclusiones muestra serias fallas metodológicas:

El primero, además de que el grupo de 25 profesores portugueses no es suficiente para conclusiones generalizables, adoleció de fallas significativas en cuanto a la calidad del pretest y a la diferencia en la preparación que se dio a los docentes del grupo control, en comparación con los del experimental. Más dudosa todavía es la generalizabilidad del segundo estudio, con un solo profesor a lo largo de 18 años. El tercer estudio, con 120 universitarios en cuatro grupos, involucró sólo a dos profesores expertos y dos novatos y analizó la frecuencia de las evaluaciones (una o tres) sin considerar en detalle el contenido y la forma de éstas.

A juicio de Dunn y Mulvenon el diseño del cuarto estudio (con 838 niños de cinco años) es bueno, pero no toma en cuenta que, además de EF, el sistema de trabajo incluía otros aspectos cuya influencia no se puede distinguir de la que haya podido tener la evaluación misma. Otros tres estudios tienen problemas similares: el quinto, en Israel, además de una muestra muy chica, se refiere a tareas que no fueron presentadas por el maestro ni se basaban en el currículo; el sexto caso sólo trabajó con alumnos de cuarto grado, con una muestra muy pequeña y con énfasis en autoevaluación; y en el séptimo caso (además de que los resultados van desde un efecto increíble de tres, hasta de sólo una desviación estándar) no se informa en qué consistieron las “discusiones generales” en el grupo control y pareciera que el grupo experimental recibió un trato distinto, más allá de lo que tenía que ver con la evaluación formativa misma.

El octavo estudio notable de Black y Wiliam, que Dunn y Mulvenon tratan en primer término, parecería en principio el más sólido, ya que se trata del meta-análisis de 21 estudios, pero incluso en este caso hay serias deficiencias: 83% de los alumnos participantes tenían necesidades educativas especiales; 72% de los efectos encontrados se presentaron en estudios que tenían “no más de dos problemas metodológicos serios”. (Dunn y Mulvenon, 2009, p. 5-7).

El texto de los profesores de Arkansas revisa nueve artículos más recientes, que se refieren en general a trabajos de educación en línea (Thompson, Goe, Paek y Ponte, 2004; Wininger, 2005; Wiliam et al., 2004; Ruiz-Primo y Furtak, 2006; Sly, 1999; Henly, 2003; Buchanan, 2000; Wang, 2007; y Velan *et al.*, 2002). Los autores reconocen que esos trabajos *ofrecen apoyo adicional a la evaluación formativa de manera fragmentada*, pero añaden que *siguen siendo problemáticos temas metodológicos similares a los de los revisados por Black y Wiliam* (Dunn y Mulvenon, 2009, p. 7). Aunque subrayan las limitaciones de los trabajos revisados, la conclusión a la que llegan no es totalmente negativa, pero sí afirman que:

> En cierta medida, las investigaciones discutidas… apoyan el impacto de la EF sobre el rendimiento de los alumnos, pero en una medida mayor apoyan la necesidad de hacer investigaciones en las que diseños y metodologías más eficientes lleven a resultados más concluyentes… no argumentamos que la EF carezca de importancia, sino sólo que la evidencia empírica que existe para apoyar "las mejores prácticas" de EF es limitada. (Dunn y Mulvenon, 2009, p. 9)

Otros trabajos recientes avanzan en la dirección señalada en el texto anterior, ya que utilizan acercamientos metodológicos más sólidos. Así lo muestra un análisis de investigaciones sobre programas de actualización para maestros en servicio que buscan mejorar sus habilidades en EF (Schneider y Randel, 2010).

Las investigaciones que revisa este capítulo se refieren a experiencias de duración considerable, en que la preparación de los participantes sobre EF implicó muchas horas, siempre se incluyó la variable relativa a los resultados de los alumnos y se manejaron números importantes de maestros, alumnos y grupos, con diseños de tipo cuasi-experimental, cuidando la comparabilidad de grupos con tratamiento y sin él, utilizando técnicas analíticas avanzadas, a falta de diseños experimentales.

El trabajo concluye con reflexiones sobre los retos metodológicos que enfrentan estos estudios, incluyendo la atención a los estándares para juzgar la calidad de las evidencias (según el repositorio *What Works Clearinghouse,)* las dificultades de manejar diseños experimentales con grupos completos (*intact classrooms*), las de las medidas de los resultados, del tamaño de las muestras, la fidelidad de implementación y la duración del estudio (Schneider y Randel, 2010, p. 267-272).

En muchos casos los resultados fueron favorables a la hipótesis de que las prácticas de EF contribuyen a mejorar el aprendizaje; en un número menor no se encontraron diferencias significativas. Nuevamente la evidencia no es concluyente, pero sí parece inclinar la balanza en el sentido de las opiniones favorables a la EF.

Otro ejemplo interesante en el sentido que se comenta es el trabajo sobre EF, motivación y aprendizaje de las ciencias naturales, de Ma. Araceli Ruiz Primo *et al.* (2010), que los autores describen como sigue:

> Un estudio de pequeñas dimensiones, aleatorizado, para someter a prueba la afirmación de Black y Wiliam (1998) de que la retroalimentación basada en EF produce un fuerte efecto positivo en el aprendizaje… el proyecto ponía a prueba una gran idea relacionada con la EF, que se podría obtener una gran ganancia en el aprendizaje con una inversión relativamente pequeña: incorporar a un currículo de ciencias utilizado en todo el país EF conceptualmente coherentes. (Ruiz Primo *et al.*, 2010, p. 143)

Después de explicar las características del estudio, cuidadosamente diseñado e implementado, los investigadores reportan los resultados relativos a los cambios esperados en los niveles de rendimiento de los alumnos como sigue:

> Sorprendentemente los resultados no corroboraron la hipótesis (…) el grupo experimental no obtuvo resultados significativamente mejores que el grupo de

> comparación ni en las pruebas de rendimiento ni en las medidas de motivación. De hecho los alumnos del grupo de comparación tuvieron resultados promedio ligeramente mejores que los del grupo experimental, aunque no estadísticamente significativos (…) la brecha entre los alumnos de alto y bajo rendimiento en el grupo experimental no fue tan grande como en el grupo de comparación (…) (Ruiz Primo *et al.,* 2010, p. 151)

La revisión de videos de clases que se grabaron permitió buscar una explicación de esos resultados, revisando la fidelidad de implementación.

El estudio de implementación buscaba entender la relación entre el tratamiento (el currículo prescrito) y las mediciones del aprendizaje (el currículo logrado), para lo cual primero se sistematizó el currículo prescrito según la guía que se había dado a los maestros participantes, y luego se analizaron las grabaciones de clases para medir en qué grado los maestros realizaron las EF como se esperaba lo hicieran. El resultado fue que había considerables diferencias en cuanto a la forma de hacer las evaluaciones, lo cual parece reflejarse en un impacto diferencial en el rendimiento. La conclusión de los investigadores es la siguiente:

> Black y Wiliam (1998) encontraron que la intervención que impacta el nivel de aprendizaje de los estudiantes es la retroalimentación. Hattie y Timperley (2007) encontraron además que la calidad de la retroalimentación impacta el grado en que ayuda a los estudiantes a mejorar. No debe sorprender que los estudiantes con mejores resultados sean los que tuvieron maestros que cierran mejor el ciclo de EF (…) usar información para ajustar enseñanza no fue bien implementado por muchos maestros del grupo experimental (…) con base en la evidencia recogida durante el estudio de implementación sabemos que, en general, los maestros conseguían que los alumnos compartieran sus ideas, pero que no conseguían utilizar la información para ajustar su propia enseñanza. Obviamente recomendar que se ajuste la enseñanza es más fácil que hacerlo. (Ruiz Primo *et al.,* 2010, p. 154)

Conviene precisar que la justificación del uso de estudios basados en diseños experimentales estrictos (que incluyan la asignación aleatoria de los sujetos a los grupos experimental y control) como soporte para llegar a conclusiones sobre el impacto de cierta intervención se justifica plenamente en principio, pues sabemos que, en ausencia de tal tipo de diseño, es problemático sacar conclusiones de tipo causal. Sin embargo, en la investigación educativa y social deben cuidarse otros aspectos para que un estudio pueda arrojar conclusiones sólidas.

En particular, es indispensable cuidar la llamada fidelidad de implementación a la que han aludido los dos últimos trabajos citados, y cuyo descuido es, al parecer, una de las deficiencias que más influyen para que los resultados de los trabajos sobre el posible impacto de la evaluación formativa no sean concluyentes. Como ha mostrado Raudenbush (2008), en la investigación sobre el efecto de ciertos fármacos sobre el organismo es sencillo garantizar que todos los sujetos de un grupo experimental reciban un tratamiento idéntico (por ejemplo “X” dosis del fármaco) y que ninguno de los sujetos del grupo control lo reciba. En educación, en cambio, y aunque se haya dado cierta preparación a los participantes, es difícil

asegurar, por ejemplo, que todos los maestros de un grupo experimental manejen prácticas de EF del mismo tipo y con idéntica calidad e intensidad, y que ninguno del grupo control utilice prácticas que puedan llevar a resultados análogos.

Por otra parte, un elemento más a tener en cuenta al estudiar el impacto de la EF tiene que ver con la dificultad de introducir prácticas novedosas, que se oponen a tradiciones muy arraigadas, como las que tienen que ver con la forma tradicional de evaluar que prevalece en las aulas desde hace muchos años. En el caso del sistema educativo mexicano, por ejemplo, si se analiza la normatividad sobre las evaluaciones que deben hacer los maestros se aprecia que casi no ha cambiado desde hace medio siglo, pese a que en ese lapso se puso de moda la pedagogía constructivista, surgieron y proliferaron las pruebas en gran escala y se comenzó a hablar de EF. Por ello no debería sorprender que muchos actores, incluyendo a maestros, pero también a alumnos y padres de familia, se sientan incómodos cuando se quiere introducir innovaciones como la que es objeto de este trabajo.

El último trabajo empírico que se revisa en esta sección tiene que ver justamente con esa resistencia, en el contexto de un sistema educativo en el que las nuevas formas de evaluación está mucho más extendido que en México.

Smith y Gorard (2005) reportan resultados de un estudio sobre las reacciones de alumnos que participaban en un proyecto que incluía la práctica de no dar calificación numéricas, como suele hacerse, para desalentar la tendencia a trabajar en función de la nota, y no por un interés intrínseco por el aprendizaje. Cuando se preguntaba a esos estudiantes cómo se sentían, las respuestas eran diversas, pero:

> (…) un número considerable de alumnos tenían opiniones bastante negativas, particularmente porque, en su opinión, el hecho de no recibir calificaciones no les permitía saber cómo orientar sus esfuerzos… cuando se preguntaba si los comentarios que recibían eran útiles, la mayoría opinaba que no les daban suficiente información para saber cómo mejorar. Tampoco pensaban que el recibir calificaciones estigmatizaría a los de bajo rendimiento… el deseo de recibir calificaciones era tan fuerte que algunos admitían que intentaban calcularlas. Esto era particularmente marcado en materias como matemáticas y lengua, en relación con las cuales los chicos admitían que sumaban las palabras bien deletreadas en pruebas de vocabulario, para calcular la calificación que habrían recibido. (Smith y Gorard, 2005, p. 31-33)

IV. Conclusión

La aplicación en el aula de los principios de la EF no es sencilla, en particular si se trata de habilidades cognitivas complejas y no de simples tareas memorísticas, ya que para ello no basta que se modifiquen las prácticas de evaluación, sino que es todo el enfoque de la enseñanza lo que debe cambiar.

El sustento teórico de la EF es sólido y el resultado de las experiencias de su aplicación permiten tener expectativas razonablemente optimistas al respecto,

pero hay también elementos que muestran que se debe proceder con cautela.

Las conclusiones de un trabajo muy reciente coinciden con las que se desprenden de esta revisión de literatura. Kingston y Nash (2011) hicieron una amplia búsqueda de textos sobre evaluación formativa y/o evaluación para el aprendizaje a partir de 1988, incluyendo revistas arbitradas o no, ponencias y tesis, en niveles educativos preuniversitarios.

ERIC permitió localizar 407 artículos y Google Scholar dio 17,300 referencias, pero la mayoría con deficiencias metodológicas tan serias que impidieron que se les considerara en el análisis. Las que reunieron los criterios necesarios para ser incluidas fueron sólo 13, en las que se encontraron 42 medidas del efecto del uso de la evaluación formativa (*effect size*). La mayor parte de estas medidas (23) se referían al efecto de programas de actualización de maestros en servicio, 7 al impacto de evaluaciones que formaban parte de los materiales curriculares, 6 se referían a evaluaciones por computadora, 3 al efecto de formas particulares de retroalimentación y 3 a aspectos de autoevaluación y coevaluación.

La mediana del tamaño del efecto fue de 0.25, menor que la de 0.7 o hasta 1.5 que se ha reportado en otros trabajos. El efecto varía de 0.09 a 0.32 dependiendo del área curricular y del tipo de intervención (Kingston y Nash, 2011, p. 32-35).

Como sugiere la parte final del título (*A call for research*), el trabajo de Kingston y Nash termina con un llamado a los interesados en el tema a realizar estudios con un buen diseño metodológico, que puedan llegar a conclusiones más sólidas y superen las limitaciones de muchos trabajos previos. Las recomendaciones de este trabajo (v. gr. evitar centrar la atención en grupos extremos, no limitarse a ver si hay efecto, sino buscar determinar cuáles son los factores que influyen en que el efecto sea mayor o menor) se añaden a las que ya se han señalado en esta revisión, en particular el uso de diseños experimentales o cuasi-experimentales y el cuidado de la fidelidad de la implementación.

Los intentos por introducir un enfoque cuya aplicación implica cambios importantes en prácticas muy arraigadas pueden ser superficiales, reduciéndose a la adopción de una terminología novedosa, sin modificar los procesos básicos de enseñanza y de aprendizaje.

La investigación deberá permitir distinguir con claridad los trabajos superficiales de los rigurosos, evitando llegar a conclusiones no matizadas que prometan resultados espectaculares de cualquier esfuerzo, por limitado que sea. Este tipo de conclusiones sin matices provoca expectativas excesivas, a las que seguirá una decepción más y el abandono de una idea realmente prometedora.

Referencias

Bangert-Drowns, R. L., Kulik, Ch., Kulik, J. A. y Morgan, M. T. (1991a). The instructional effect of feedback on test-like events. *Review of Educational Research, 61*(2), 213-238.

Bangert-Drowns, R. L., Kulik, J. A. y Kulik, Ch. (1991b). Effects of frequent classroom testing. *Journal of Educational Research, 85*, 89-99.

Black, P. J. (1993). Formative and summative assessment by teachers. *Studies in Science Education, 21*, 49-97.

Black, P. y Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7-74.

Bloom, B. S. (1984a). The search for methods of group instruction as effective as one-to-one tutoring. *Educational Leadership, 41*(8), 4-17.

Bloom, B. S. (1984b). The 2-sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*(6), 4-16.

Buchanan, T. (2000). The efficacy of a World-Wide Web mediated formative assessment. *Journal of Computer Assisted Learning, 16*, 193-200.

Centre for Educational Research and Innovation (2005). *Formative assessment. Improving learning in secondary classrooms*. París: OECD.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*, 438-481.

Dunn, K. E. y Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment Research and Evaluation, 14*(7), 1-11.

Guskey, T. R. (2007). Formative classroom sssessment and Benjamin S. Bloom: Theory, research and practice. En J. H. McMillan, *Formative classroom assessment*: *Theory* into *practice* (63-78). Nueva York: Teachers College Press.

Hattie, J. (1992). Measuring the effects of schooling. *Australian Journal of Education, 36*(1), 5-13.

Hattie, J. y H. Timperley (2007). The power of feedback. *Review of Education Research, 77*(1), 81-112.

Henly, D. C. (2003). Use of Web-based formative assessment to support student learning in a metabolism/nutrition unit. *European Journal of Dental Education, 7*, 116-122.

Kingston, N. y Brooke, N. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28-37.

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

Kulik, C. C., Kulik, J. A. y Bangert-Drowns, R. L. (1990). Effectiveness of mastery learning programs: A meta-analysis. *Review of Educational Research, 60*(2), 265-299.

Marzano, R. (2007). Designing a comprehensive approach to classroom assessment, en D. Reeves (Ed.) *Ahead of the curve* (pp. pp. 103-125), Bloomington: Solution Tree Press.

Meisels, S., Atkins-Burnett, S., Xue, Y., DiPrima, D. y Son, S. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement scores. *Educational Policy Analysis Archives, 11*(9).

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22*, 155-175.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal, 45*(1), 206-230.

Rodríguez, M. C. (2004). The role of classroom assessment in pupil performance in TIMSS. *Applied Measurement in Education, 17*(1), 1-24.

Ruiz-Primo, M. A. y Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning. *Educational Assessment, 11*(3-4), 205-235.

Scheneider, M. C. y Randel, B. (2010). Research on characteristics of effective professional development programs for enhancing educators' skills in formative assessment. En Andrade y Cizek (Eds.), *Handbook of formative assessment* (pp. 251-276). Nueva York-Londres: Routledge.

Sly, L. (1999). Practice tests as formative assessment improve student performance on computer managed learning assessments. *Assessment and Evaluation in Higher Education, 24*(3), 339-343.

Smith, E. y Gorard, S. (2005). They dont't give us our marks: the role of formative feedback in student progress. *Assesment in Education: principles, policy & practice, 12*(1), 21-38.

Stiggins, R. J. (2007). Conquering the formative assessment frontier. En J. H. McMillan (Ed.) *Formative classroom assessment: Theory into practice* (pp. 8-27). Nueva York: Teachers College Press.

Stiggins, R. J. (2001). Unfulfilled promise of classroom assessment. *Educational Measurement: Issues & Practice, 20*(3), 5-15.

Stiggins, R. J. y Arter, J. (2002). Assessment for learning. International Perspectives. The Proceedings of an International Conference. Documento presentado en la *Annual Meeting of the National Council on Educational Measurement*, Nueva Orleans.

Thompson, M., Goe, L., Paek, P. y Ponte, E. (2004). *Study of the California formative assessment and support system for teachers: Relationship of BTSA/CFASST and student achievement*. Princeton: Educational Testing Service.

Velan, G. M., Rakesh, K. K., Mark, D. y Wakefield, D. (2002). Web-based self-assessments in Pathology with questionmark perception. *Pathology, 34*, 282-284.

Walberg, H. J. (1984). Improving the productivity of America's schools. *Educational Leadership, 41*(8), 19-27.

Wang, T. H. (2007). What strategies are effective for formative assessment in an e-learning environment? *Journal of Computer Assisted Learning, 23*, 171-186.

Wiliam, D., Lee, C., Harrison, C. y Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education, 11*, 49-65.

Wininger, R. S. (2005). Using your tests to teach: Formative summative assessment. *Teaching Psychology, 32*(2), 164-166.