Formative Assessment: Can the Claims for Effectiveness Be Substantiated?[1]

Randy Elliot Bennett
Educational Testing Service
Princeton, New Jersey
USA

[1] This paper is adapted from R. E. Bennett, *A critical look at the meaning and basis of formative assessment* (RM-09-06). Princeton, NJ: Educational Testing Service, 2009.

July 2009

Abstract

In primary and secondary education, formative assessment has become a common theme at conferences, a standard offering in test company catalogues, and a focus for teacher in-service training. A key reason for the popularity of formative assessment is, undoubtedly, the claims that have been made for its effectiveness. The most commonly cited quantitative claims are for effects between .4 and .7 standard deviations, although some individuals have asserted effects of up to 2 standard deviations. This paper reviews the evidentiary sources cited for these claims, and suggests how the effectiveness of formative assessment might be represented more responsibly.

In primary and secondary education, formative assessment is in vogue. It has become a key topic at educational conferences, a standard offering in virtually every test publisher's catalog, and a frequent focus for teacher in-service training. The concept has been capitalized upon by test publishers (Popham, 2006; Shepard, 2008) and educational consultants alike to the point that some observers think it, too, may be just another passing fad. Much of the interest in formative assessment is based upon very strong claims for effectiveness, which is the primary focus of this paper.

Claims for Effectiveness

The most widely cited source for these strong claims is almost certainly the pair of 1998 papers published by Paul Black and Dylan Wiliam. "Inside the Black Box" is a brief position piece that appeared in *Phi Delta Kappan* (Black & Wiliam, 1998a). That article summarizes a lengthy, meticulous review, "Assessment and Classroom Learning," published the same year in the journal *Assessment in Education* (Black & Wiliam, 1998b).

As noted, one or the other of the articles has been used routinely to undergird claims for the effectiveness of formative assessment. For example, one highly respected testing expert wrote:

> Based on their meta-analysis, Black and Wiliam [1998] report effect sizes of between .4 and .7 in favor of students taught in classrooms where formative assessment was employed. (Popham, 2008, p. 19)

A second, respected expert stated:

> English researchers Paul Black and Dylan Wiliam recently published the results of a comprehensive meta-analysis and synthesis of more than 40 controlled studies of the impact of improved classroom assessment on student success ... (Stiggins, 1999)

This author then cited the same .4 to .7 effect sizes as claimed above, a gain that is roughly double the average growth US children in the upper primary to lower secondary grades would be expected to make on standardized tests in a school year.[2]

In a later article, the same author appeared to expand the claim, both in terms of the magnitude of the observed effects and the size of the evidence base:

> Black and Wiliam, in their 1998 watershed research review of more than 250 studies from around the world on the effect of classroom assessment, report gains of a half to a full standard deviation. (Stiggins, 2006, p. 15)

On the same page, the effectiveness claim appeared to be made stronger still:

[2] Expected growth was calculated from the norms of the *Metropolitan Achievement Test Eighth Edition* (Harcourt Educational Measurement, 2002), the *Iowa Tests of Basic Skills Complete Battery* (Hoover, Dunbar, & Frisbie, 2001), and the *Stanford Achievement Test Series Tenth Edition* (Pearson, 2004).

> Bloom and his students (1984) made extensive use of classroom assessment … for learning … [and] reported subsequent gains in student test performance of one to two standard deviations. (Stiggins, 2006, p. 15)

Similar claims are found in a conference presentation by a third expert, available on the World Wide Web (Kahl, 2007). Here, the expanded effects of a half to a full standard deviation previously attributed to Black and Wiliam by the second expert are repeated. This expansion is consequential because, in the most commonly cited effect-size classification (Cohen, 1988), effects that previously would have been considered small-to-medium have now become medium-to-large.[3]

In addition to referencing work by Black and Wiliam and by Bloom, Kahl's presentation (2007) cites Meisels, Atkins-Burnett, Xue, and Bickel (2003) and Rodriguez (2004). These studies have also been used by many others as supportive evidence (e.g., Arter, 2006, p. 42; Davies, n.d.; Glasson, 2008; Love, 2009, p. 15; Stiggins, 2006).

Effect sizes are not the only metric in which impact claims have been made. The same basic assertions appear phrased in terms of improving student performance a given number of percentile points in the achievement distribution, increasing student learning by some number of months or years, or even moving countries who performed middling on international assessments like PISA or TIMSS to the top of the pack (e.g., Chappuis, Chappuis, & Stiggins, 2009, p. 56).

A Brief Review of the Evidence

Regardless of the metric used, the essential argument put forth by these and numerous other advocates is that empirical research proves formative assessment causes medium-to-very large achievement gains and that these results come from trustworthy sources. In particular, the sources are said to include meta-analyses, as well as noteworthy individual studies.

These claims deserve a closer look. The idea of meta-analysis is a sensible place to start because it has been so frequently cited in the effectiveness claims to connote methodological rigor. Meta-analysis was originally conceived of as a method for describing the empirical results observed in a research literature. In its simplest form, the method is essentially a pooling of results from a set of comparable studies that yields one or more summary statistics, including what is commonly called an *effect size*. For experimental studies, the effect size is typically computed as the difference between the treatment-group and control-group means, divided by the standard deviation (of the treatment group or pooled across the groups).

Like any method, however, meta-analysis can produce meaningless results. The results should be considered suspect when, for example:

- studies are too disparate in topic to make summarization meaningful;
- multiple effects too often come from the same study or from studies authored by the same individuals, and no accounting for such violations of independence has been made;
- study characteristics, such as technical quality or datedness, are not considered; or
- the meta-analysis itself is not published so that the methods involved are unavailable for critical review.

[3] Cohen (1988, pp. 25-27) considered effects of .2 to be small, .5 to be medium, and .8 to be large.

In this regard, a major concern with the original Black and Wiliam (1998a) review is that the research covered is too disparate to be summarized meaningfully through meta-analysis. That research includes studies related to feedback, student goal orientation, self-perception, peer assessment, self-assessment, teacher choice of assessment task, teacher questioning behavior, teacher use of tests, and mastery learning systems. That collection is simply too diverse to be sensibly combined and summarized by a single effect-size statistic.

This fact might be better appreciated if more advocates of formative assessment carefully read the original article. In a section titled, "No Meta Analysis," Black and Wiliam (1998b) stated the following:

> It might be seen desirable… for a review of this type to attempt a meta-analysis of the quantitative studies that have been reported… Individual quantitative studies which look at formative assessment as a whole do exist..., although the number with adequate and comparable quantitative rigour would be of the order of 20 at most. However, whilst these [studies] are rigourous within their own frameworks and purposes, … the underlying differences between the studies are such that *any amalgamations of their results would have little meaning*. (p. 53) (emphasis added)

In their review article, then, Black and Wiliam, report no meta-analysis of their own doing, nor any quantitative results of their own making. The confusion may occur because, in their brief *Phi Delta Kappan* position paper, Black and Wiliam (1998a) do, in fact, attribute a range of effect sizes to formative assessment. However, no source for those values is ever given. As such, these effect sizes are not the quantitative result, meta-analytical or otherwise, of the 1998 *Assessment in Education* review but, rather, a misinterpretation that has arguably become the educational equivalent of urban legend.[4] Even so, the review provides a very valuable qualitative synthesis, though of a broad array of literatures and not of a single, well-defined class of treatments that could be called formative assessment.

Whereas the Black and Wiliam articles are probably the most frequent derivation for the claimed large impact of formative assessment, as suggested earlier there are a number of other commonly referenced sources. But each source raises concerns that might call the size of the claimed effects into question.

Let's start with the Bloom studies that reputedly found effects of between 1 and 2 standard deviations, somewhere between *large* and *huge*. That claim comes from a summary article (Bloom, 1984), based primarily on (now quite dated) dissertations conducted by Bloom's students. In a comprehensive literature review that included those same studies, Slavin (1987, p. 207) wrote:

> Bloom's claim that mastery learning can improve achievement by more than 1 sigma is based on brief, small, artificial studies that provided additional instructional time to the experimental classes [and not to controls]. In longer term and larger studies with experimenter-made measures, effects of group-based mastery learning are much closer to

[4] It is possible that these values represent Black and Wiliam's retrospective extraction from the 1998 review of the range of mean effects found across multiple meta-analytical studies done by other investigators on different topics (i.e., the mean effect found in a meta-analysis on one topic was .4 and the mean effect found in a meta-analysis on a second topic was .7). If so, the range of observed individual study effects would, in fact, be wider than the oft-quoted .4 to .7, as each mean itself represents a distribution. But more fundamentally, the construction of any such range would seem subject to Black and Wiliam's (1998b) very own critique--i.e., "… the underlying differences between the studies are such that any amalgamations of their results would have little meaning" (p. 53).

> 1/4 sigma, and in studies with standardized measures there is no indication of any positive effect at all. [The] 1-sigma claim is misleading … and potentially damaging … as it may lead researchers to belittle true, replicable, and generalizable achievement effects in the more realistic range of 20-50% of [a] standard deviation.

A second commonly referenced source, by Nyquist (2003), is far more recent. The relevance of this source to the primary and secondary school context can be immediately questioned because, although rarely noted in advocates' invocations, it focuses on the college-level population (p. 19). Second, the study is an unpublished master's thesis and, as such, is not generally available. The fact that it is unpublished lessens its value as backing for the general efficacy of formative assessment since it has not been subjected to peer review—a hallmark of the scientific process—nor has it been readily accessible for purposes of challenge from the field and rejoinder by the author.

Two individual studies, by Meisels et al. (2003) and by Rodriguez (2004), also have figured among advocates' evidentiary sources. Of note is that both studies were observational, so it is not possible to rule out alternative explanations for treatment effects. The design of the Meisels et al. study is of particular concern since it seems to have used a volunteer treatment group (ostensibly more motivated than the comparison group) and because other curricular innovations were being implemented during the study period. No accounting was apparently made for either the potential selection bias or the confound with other innovations, so defensible assertions about the impact of formative assessment are very difficult to make.

In keeping with the nature of an observational investigation, Rodriguez (2004) was quite modest in his claims. The study's analysis was complicated, incorporating many variables, with no clear interpretation possible regarding a cause-and-effect relationship between formative assessment and student achievement. The study did report achievement effects related to classroom self-efficacy and to uncontrollable attributions, constructs that might be theoretically linked to formative assessment practice (e.g., Stiggins, 2006). But how these variables are connected directly in the study to classroom (formative) assessment is not evident, nor is the direction of their causal relationship to achievement. The only variable that might be considered to directly represent classroom assessment practice is the use of teacher-made tests, for which the effect on achievement (controlling for all other model variables) was *negative* (i.e., the more use of classroom testing, the lower the achievement). Given these facts, it is very difficult to see how this study legitimately supports efficacy claims.

The last source to be mentioned is that of Kluger and DeNisi (1996). This article included a (real) meta-analysis of a large number of studies. The article was published in a very high-quality journal—*Psychological Bulletin*—and was focused on one topic relevant to formative assessment (i.e., feedback). In that sense, the analysis was far more focused than the very broad range of the Black and Wiliam (1998a) review. All the same, the Kluger and DeNisi analysis included a wide variety of criterion measures spanning academic and employment contexts (e.g., reading errors, arithmetic computation, test performance, memory retention, reaction time, puzzles).

With respect to results, of special note is that Kluger and DeNisi found a mean effect size for the impact of feedback on performance of .41, less than the much larger effects often claimed for formative assessment. These investigators also found that 38% of the effects were *negative,* meaning that the control condition was more effective than whatever constituted the feedback intervention in well over a third of cases. Finally, feedback appeared to improve performance far more dramatically on simple versus complex tasks and had no impact on transfer, leading Kluger

and DeNisi to conclude that, "… the evidence for any learning effect here was minimal at best" (p. 278).[5]

Improving Our Claims

In short, then, the research does not appear to be as unequivocally supportive of formative assessment practice as it is sometimes made to sound. Given that fact, how might we improve the quality of the claims we make for the efficacy of formative assessment? An obvious first step should be exercising greater care in the evaluation of sources of evidence and in the attributions we make about them. Second, a clearer definition of what we mean by formative assessment is essential to helping abstract a class of things to study and make claims about. Current "process" definitions (e.g., Black & Wiliam, 1998a; p. 140, McManus, 2008, p. 3; Popham, 2008, p.6) are inadequate because they are too broad to allow either consistent implementation or meaningful analysis of effectiveness.

A stronger definition would arguably include (1) a conceptual framework that identifies the characteristics and components of the thing we are claiming is "formative assessment," along with the rationale for each of those characteristics; (2) a theory of action that postulates how these characteristics and components work together to create some desired outcome; and (3) a concrete instantiation that illustrates what it looks like and how it might work in a real setting.

In this respect, the theory of action is particularly important because, without it, we can't meaningfully evaluate the underlying mechanisms that are supposed to cause the intended effects. Among other things, a theory of action for formative assessment would seem to require that practitioners carefully design situations (or ask questions) to elicit evidence on critical components of domain understanding; make inferences from that evidence about what students know and can do; and use those inferences as a basis for adapting instruction.

Focusing on these action-theory mechanisms suggests that, to be considered effective, formative assessment requires at least two types of argument: a *validity argument* to support the quality of inferences about students and an *efficacy argument* to support the impact on learning and instruction.

Each argument requires backing, both logical and empirical. The validity argument makes claims about the meaning of the evidence elicited through formative assessment (e.g., that a student needs instruction in a particular reading-skill component). Backing to support those claims might include data showing that different observers draw similar inferences about a student's skills from the same evidence; that the inferences drawn are consistent with other, more in-depth methods of characterizing what a student knows and can do (e.g., with a very carefully constructed, targeted assessment of a particular skill component or with information from a variety of other sources); and that different observers make substantively similar adjustments to instruction from the same evidence.

In contrast to the validity argument, the efficacy argument makes claims about changes in student skill associated with the use of formative assessment. This argument presumes that those

[5] Advocates sometimes interpret the feedback research to indicate that the positive results are attributable to practices consistent with formative assessment (e.g., qualitative characterizations drawing attention to task performance) and the negative results to antithetical practices (e.g., numeric or letter grades drawing attention to self). Although there is certainly some support for this position in the Kluger and DeNisi (1996) results, it would appear to be an oversimplification. For example, Kluger and DeNisi noted that feedback effects "… are moderated by the nature of the task … [and] the exact task properties that moderate [these] effects are still poorly understood" (p. 275). In a more recent review, Shute (2008) wrote, "Within this large body of feedback research, there are many conflicting findings and no consistent pattern of results" (p. 153).

changes are caused by actions the teacher (or student) takes based on assessment inferences. (Otherwise why bother to observe and judge student behavior?) Thus, logically, the efficacy argument for formative assessment must include the validity argument.

Aside from the backing to support the validity argument, backing for the efficacy argument might include data related to several areas. First, that backing should include data showing that the formative assessment was implemented as intended. Second, that backing should include data suggesting that other intermediate outcomes stipulated by the theory of action were achieved. Finally, the backing should entail data indicating that students participating in formative assessment changed more in a positive direction on ultimate outcomes of interest than those participating in some alternative practice (e.g., that students do, in fact, act on feedback, become more engaged, and learn more). In short, the efficacy argument provides the backing not only for claims that formative assessment is effective but, as importantly, for whether those effects are caused by the practices composing the action theory.

Conclusion

The term *formative assessment* does not yet represent a well-defined set of artifacts or practices. A stronger definition would arguably include a conceptual framework, a theory of action, and one or more concrete instantiations. Such a definition would allow for more consistent implementation and more meaningful evaluation of effectiveness, which current conceptual definitions do not.

The primary and secondary school effectiveness research does suggest that the practices associated with formative assessment can, under the right conditions, facilitate learning. However, these effects may vary markedly across implementations of the multiplicity of practices that fall under current definitions of formative assessment, as well as across subpopulations of students. (Consider the variation in the effectiveness of feedback as one example.) Also, quantitative claims for the efficacy of formative assessment should be viewed with caution. The most frequently cited effect-size claim of .4 - .7 standard deviations is neither meaningful as a representation of the impact of a single well-defined class of treatments, nor readily traceable to *any* inspectible, empirical source. Other empirical sources are dated, unpublished, methodologically flawed, target older populations, or show smaller effects than proponents cite. Finally, the validity argument, and backing for it—both key to an action theory of formative assessment—are generally absent. Given these facts, researchers might be more responsible in their claims and educators less immediately accepting of those who push too self-assuredly for large-scale adoption of formative assessment.

References

Arter, J. (2006). Making use of data: What educators need to know and be able to do. In J. O'Reilly (Ed.), *Beyond NCLB: From measuring status to informing improvement* (pp. 39–72). Retrieved February 25, 2009 from http://natd.org/files/uplink/2006proceedings.pdf

Black, P., & Wiliam, D. (1998a). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Black, P., & Wiliam, D. (1998b). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.

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

Chappuis, J., Chappuis, S., & Stiggins, R. (2009). Formative assessment and assessment for learning. In L. M. Pinkus (Ed.), *Meaningful measurement: The role of assessments in improving high school education in the Twenty-First Century*. Washington, DC: Alliance for Excellent Education. Retrieved August 3, 2009 from http://www.all4ed.org/files/MeanMeasCh3ChappuisStiggins.pdf

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Davies, A. (n.d.). Summary of classroom assessment research. *Assessment for learning: An online resource for educators*. Retrieved February 25, 2009, from http://annedavies.com/assessment_for_learning_arc.html

Glasson, T. (2008). Improving student achievement through assessment for learning. *Curriculum Leadership, 6*(31). Retrieved February 25, 2009, from http://www.curriculum.edu.au/leader/improving_student_achievement,25374.html?issueID=11603

Harcourt Educational Measurement. (2002). *Metropolitan8: Technical manual*. San Antonio, TX: Author.

Hoover, H. D., Dunbar, S. B., & Frisbie, D A. (2001). *Iowa Tests of Basic Skills Complete/Core Battery: Spring norms and score conversions with technical information*. Itasca, IL: Riverside.

Kahl, S. (2007, April 23). *Formative assessment: An overview* (PowerPoint presentation). Retrieved February 11, 2009, from the Montana Office of Public Instruction Web site: http://www.opi.state.mt.us/pdf/Assessment/conf/Presentations/07MON_FormAssmt.ppt

Kluger, A. N., & 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.

Love, N. (2009). Building a high-performance data culture. In N. Love (Ed.), *Using data to improve learning for all: A collaborative inquiry approach* (pp. 2–24). Thousand Oaks, CA: Corwin Press.

McManus, S. (2008). *Attributes of effective formative assessment*. Washington, DC: Council for Chief State School Officers. Retrieved March 2, 2009 from http://www.ccsso.org/publications/details.cfm?PublicationID=362

Meisels, S. J., Atkins-Burnett, S., Xue, Y., Bickel, D. D., & Son, S. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. *Education Policy Analysis Archives, 11*(9). Retrieved Februrary 11, 2009, from http://epaa.asu.edu/epaa/v11n9/

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

Pearson. (2004). *Stanford Achievement Test Series Tenth Edition: Technical data report*. Iowa City, IA: Author.

Popham, W. J. (2006). Phony formative assessments: Buyer beware! *Educational Leadership, 64*(3), 86–87.

Popham, W. J. (2008). *Transformative assessment*. Alexandria, VA: ASCD.

Rodriguez, M. C. (2004). The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education, 17,* 1–24.

Shepard, L. A. (2008). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279–303). New York: Erlbaum.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78,* 153–189.

Slavin, R. E. (1987). Mastery learning reconsidered. *Review of Educational Research, 57,* 175–213.

Stiggins, R. J. (1999). Assessment, student confidence, and school success**.** *Phi Delta Kappan, 81*(3), 191–198.

Stiggins, R. (2006). Assessment for learning: A key to motivation and achievement. *Edge, 2*(2), 3–19.