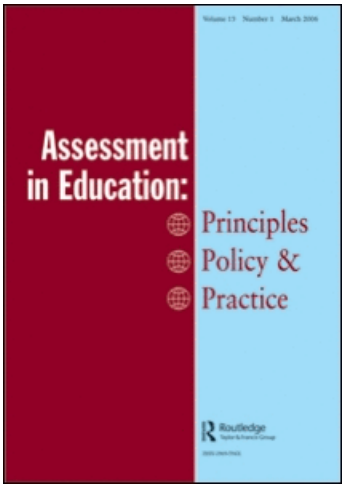

Assessment in Education: Principles, Policy & Practice
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713404048

Formative assessment: a critical review

Randy Elliot Bennett[a]
[a] Research and Development, Educational Testing Service, Princeton, NJ, USA

Online publication date: 25 January 2011

To cite this Article Bennett, Randy Elliot(2011) 'Formative assessment: a critical review', Assessment in Education: Principles, Policy & Practice, 18: 1, 5 — 25
To link to this Article: DOI: 10.1080/0969594X.2010.513678
URL: http://dx.doi.org/10.1080/0969594X.2010.513678

Assessment in Education: Principles, Policy & Practice
Vol. 18, No. 1, February 2011, 5–25

Routledge
Taylor & Francis Group

Formative assessment: a critical review

Randy Elliot Bennett*

Research and Development, Educational Testing Service, Princeton, NJ, USA

This paper covers six interrelated issues in formative assessment (aka, 'assessment for learning'). The issues concern the definition of formative assessment, the claims commonly made for its effectiveness, the limited attention given to domain considerations in its conceptualisation, the under-representation of measurement principles in that conceptualisation, the teacher-support demands formative assessment entails, and the impact of the larger educational system. The paper concludes that the term, 'formative assessment', does not yet represent a well-defined set of artefacts or practices. Although research suggests that the general practices associated with formative assessment can facilitate learning, existing definitions admit such a wide variety of implementations that effects should be expected to vary widely from one implementation and student population to the next. In addition, the magnitude of commonly made quantitative claims for effectiveness is suspect, deriving from untraceable, flawed, dated, or unpublished sources. To realise maximum benefit from formative assessment, new development should focus on conceptualising well-specified approaches built around process and methodology rooted within specific content domains. Those conceptualisations should incorporate fundamental measurement principles that encourage teachers and students to recognise the inferential nature of assessment. The conceptualisations should also allow for the substantial time and professional support needed if the vast majority of teachers are to become proficient users of formative assessment. Finally, for greatest benefit, formative approaches should be conceptualised as part of a comprehensive system in which all components work together to facilitate learning.

Keywords: formative assessment; assessment for learning

In primary and secondary education, formative assessment is, without doubt, in vogue. It has become a common theme at educational conferences, a standard offering in test-company catalogues, the subject of government tenders, and a focus for teacher in-service training.

This paper examines six interrelated topics, denoted as follows: the definitional issue, the effectiveness issue, the domain dependency issue, the measurement issue, the professional development issue, and the system issue. Collectively, these topics are important in understanding what formative assessment is and what claims we, as responsible professionals, should be making about it. The purpose of the paper is to encourage something largely missing from the discourse around formative assessment today: that is, a frank and judicious dialogue, one that is necessary for moving this promising concept forward.

*Email: rbennett@ets.org

ISSN 0969-594X print/ISSN 1465-329X online

DOI: 10.1080/0969594X.2010.513678
http://www.informaworld.com

The definitional issue

What, exactly, is 'formative assessment'? The distinction between the summative and formative roles was first proposed by Scriven (1967) in the context of programme evaluation (Black and Wiliam 2003; Wiliam and Thompson 2008). For Scriven, summative evaluation provided information to judge the overall value of an educational programme (as compared with some alternative), whereas the results of formative evaluation were targeted at facilitating programme improvement. It was Bloom (1969) who, using the very same terminology, made a similar distinction, but with respect to students (Black and Wiliam 2003; Wiliam and Thompson 2008). For Bloom (1969, 48), the purpose of formative evaluation was '… to provide feedback and correctives at each stage in the teaching-learning process'. Summative evaluation was employed to judge what the learner had achieved at the end of a course or programme.

Over the years, much work has been directed at elaborating Bloom's distinction, especially in Australia (e.g., Sadler 1989) and in the United Kingdom through the Assessment Reform Group.[1] Even so, the essence of Bloom's distinction holds today, although the term 'formative assessment' has been substituted to connote a focus on students instead of programmes.

More interesting from a definitional perspective is that, with its recent 'return' to US education from abroad, the concept has become somewhat confused. According to the September 17, 2007 edition of *EdWeek*, the 'Test industry [is] split over "formative" assessment', so much so that 'Testing expert Richard J. Stiggins … has stopped using the term …' (Cech 2007, 1).

The 'split' referenced by *EdWeek* has on one side those who believe 'formative assessment' refers to an instrument (e.g., Pearson 2005), as in a diagnostic test, an 'interim' assessment, or an item bank from which teachers might create those tests. Formative assessments of this kind will typically produce one or more scores, often claimed to have 'diagnostic' value, and will generally require cycle times suited more to instructional units and marking periods than to daily lessons (Wiliam and Thompson 2008). Not surprisingly, this view of formative assessment is quite common among test publishers; it represents something they understand and can provide. Examples include the Pearson *Progress Assessment Series*, CTB/McGraw-Hill's *Accuity* (http://www.acuityforschool.com/), and the *ETS Formative Assessment Item Bank*.

The other side of this split – populated more by educators and researchers than test publishers – is the view that '… formative assessment is not a test but a process…' (Popham 2008, 6). In this view, the process produces not so much a score as a qualitative insight into student understanding (Shepard 2008). This camp further argues that *the* distinguishing characteristic is ' … when the [results are] actually used to adapt the teaching to meet student needs' (Black and Wiliam 1998a, 140). Such adaptation will typically occur over short cycles, within or between lessons (Wiliam and Thompson 2008). These ideas are brought together in a recent definition promulgated by the Formative Assessment for Teachers and Students (FAST) State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State School Officers (CCSSO) and contained in McManus (2008, 3): 'Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes'. A common simplification of this position is that as long as the results are used to change instruction, *any* instrument may be used formatively,

regardless of its original intended purpose (e.g., Weeden, Winter, and Broadfoot 2002, 28; Wiliam and Thompson 2008, 71).

Arguably, each position is an oversimplification. It is an oversimplification to define formative assessment as an instrument because even the most carefully constructed, scientifically supported instrument is unlikely to be effective instructionally if the process surrounding its use is flawed. Similarly, it is an oversimplification to define formative assessment as a process since even the most carefully constructed process is unlikely to work if the 'instrumentation', or methodology, being used in that process is not well-suited for the intended purpose. 'Process' cannot somehow rescue unsuitable instrumentation, nor can instrumentation save an unsuitable process. A strong conceptualisation needs to give careful attention to each component, as well as to how the two components work together to provide useful feedback.

As noted, *EdWeek* reported that Richard Stiggins has stopped using the term 'formative assessment', presumably because that phraseology had lost its meaning. Many advocates of the process view appear to prefer, 'assessment *for* learning', employing 'assessment *of* learning' to denote 'summative assessment'. From a definitional perspective, however, this substitution is potentially problematic in that it absolves summative assessment from any responsibility for supporting learning. Further, it, too, potentially leads to oversimplifying what is, in fact, a more complex relationship.

The relationship is more complex because a summative assessment should fulfil its primary purpose of documenting what students know and can do but, if carefully crafted, should also successfully meet a secondary purpose of support *for* learning. Such support may be provided in at least three ways. First, if the content, format and design of the test offer a sufficiently rich domain representation, preparing for the summative test can be a valuable learning experience (Shepard 2006). When students are motivated to prepare, studying encourages consolidation and organisation of knowledge, rehearsal of domain-relevant processes and strategies, stronger links to conditions of use, and greater automaticity in execution; in other words, the development of expertise. Second, recent research suggests that taking a test can both enhance learning by strengthening the representation of information retrieved during the test and also slow the rate of forgetting (Rohrer and Pashler 2010). A final way in which summative assessment may support learning is by providing a limited type of formative information. There is no claim that just *any* summative assessment can support learning effectively; only those summative tests that are designed to fulfil that subsidiary purpose by, for example, linking performance to theoretically or empirically based learning progressions (Corcoran, Mosher, and Rogat 2009; Popham 2008, 23), or mapping key tasks to the score scale (Zwick et al. 2001). With such a test, the teacher may be able to identify particular students needing more focused formative follow-up or content that may need to be re-taught presently, or taught differently next cycle.

By the same token, well-designed and implemented formative assessment should be able to suggest how instruction should be modified, as well as suggest impressionistically to the teacher what students know and can do. Thus, we should be able to design assessment systems in which summative tests, besides fulfilling their primary purposes, routinely advance learning, and formative assessments routinely add to the teacher's overall informal judgments *of* student achievement (see Table 1).

Formative assessment then might be best conceived as neither a test nor a process, but some thoughtful integration of process *and* purposefully designed methodology or instrumentation. Also, calling formative assessment by another name may only exacerbate, rather than resolve, the definitional issue.[2]

Table 1. A more nuanced view of the relationship between assessment purpose and assessment type.

	Purpose	
Type	Assessment *Of* Learning	Assessment *For* Learning
Summative	*X*	*x*
Formative	*x*	*X*

Note: X = primary purpose; x = secondary purpose.

But why is definition important in the first place? Definition is important because if we can't clearly define an innovation, we can't meaningfully document its effectiveness. Part of that documentation needs to be an evaluation of whether the formative assessment was implemented as intended, which we cannot accomplish if we don't know what was supposed to be implemented. Similarly, if we can't clearly define an innovation, we can't meaningfully summarise results across studies because we won't know which instances to include in our summary. Last, we won't be able to transport it to our own context, for how will we know the characteristics on which to focus in doing the transport?

For a meaningful definition of formative assessment, we need at least two things: a theory of action and a concrete instantiation. Among other things, the theory of action: (1) identifies the characteristics and components of the entity we are claiming is 'formative assessment', along with the rationale for each of those characteristics and components; and (2) postulates how these characteristics and components work together to create some desired set of outcomes (Bennett 2010). The concrete instantiation illustrates what formative assessment built to the theory looks like and how it might work in a real setting.

In this regard, the *Keeping Learning on Track*® Program (ETS 2010), or *KLT*, is a provocative example because it contains a rudimentary theory of action and a concrete instance to illustrate at least one type of 'formative assessment'. The theory of action revolves around 'one big idea and five key strategies', based in substantial part on the work of Black and Wiliam (1998c, 2009). The big idea is of 'students and teachers using evidence … to adapt teaching and learning to meet immediate learning needs minute-by-minute and day-by-day' (ETS 2010).

The five key strategies are Sharing Learning Expectations (i.e., clarifying and sharing learning intentions and criteria for success), Questioning (i.e., engineering effective classroom discussions, questions and learning tasks that elicit evidence of learning), Feedback, Self Assessment (i.e., activating students as the owners of their own learning), and Peer Assessment (i.e., activating students as instructional resources for one another). These strategies are used to direct the instructional processes of establishing where learners are (e.g., through questioning), where they are going (by sharing learning expectations), and how to get them there (through feedback) (Wiliam and Thompson 2008). The *KLT* strategies are implemented through teacher- and student-use of a large catalogue of techniques, including ones like, 'Three stars and a wish', and 'Traffic lights'. In 'Three stars …', students exchange work and each student is expected to indicate three things he or she liked about his or her peer's work and one thing that he or she wished could be made better. In 'Traffic lights …', each student is given a red, a yellow, and a green cup, and asked to display at key points in

the class lesson the cup indicating his or her level of understanding (i.e., red for 'don't understand', yellow for 'unsure', and green for 'please proceed').

Anticipating an issue to be discussed later, it is worth pointing out that the five key strategies are intended to be general, domain-independent ones. The strategies have links to cognitive-scientific theory, particularly that segment of the field concerned with learning through social interaction (e.g., Vygotsky 1978, as cited in Shepard 2006). 'Sociocultural' theories postulate that students learn most effectively through interchange with others, especially more proficient domain practitioners who can model the internal standards and habits of mind that define advanced competency. Sharing expectations, questioning, feedback, self-assessment, and peer assessment are intended to, among other things, help students develop internal standards for their work, reflect upon it, and take ownership of learning.

A logic model depicting the *KLT* theory of action is shown in Figure 1 (ETS 2009). The model is read from left to right. In broad strokes, the *KLT* components are postulated to cause change in teacher practice (shown in the centre area) that, in turn, influences student behaviour and increases achievement (shown in the right-hand portion).

The *KLT* components are focused on training teachers in formative assessment. The components include both materials and facilitated events. One of the events is a workshop for school staff members who will, in turn, support local teachers by helping them establish a 'learning community' for themselves. The role of the learning community

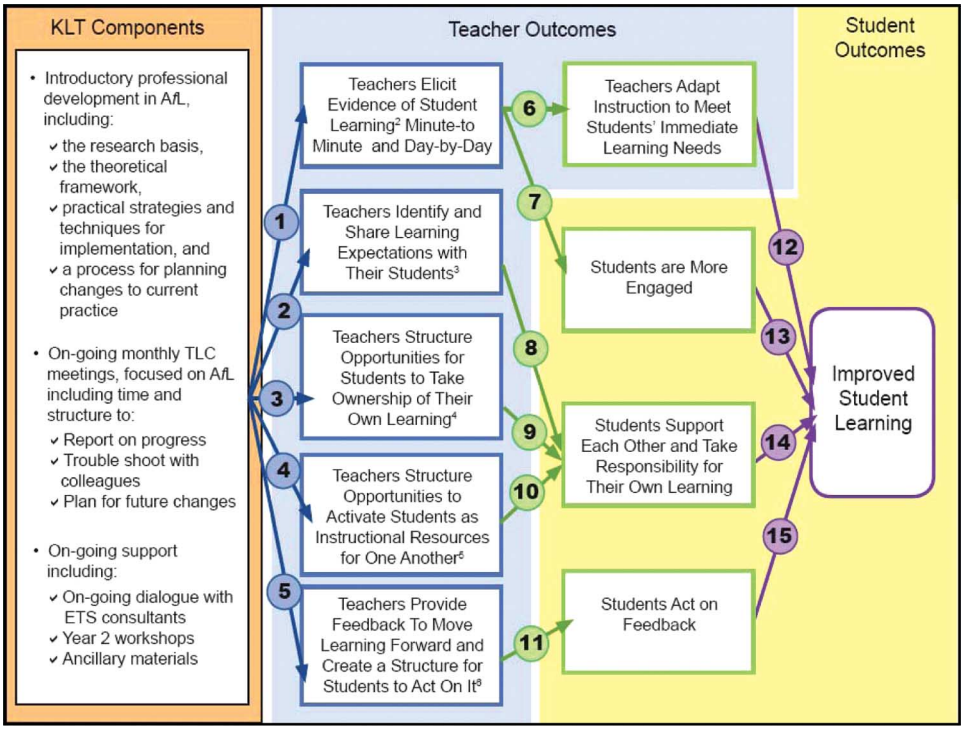

Figure 1. The *KLT* theory of action.
Note: From ETS (2009). Used by permission. © Educational Testing Service. All rights reserved.

is to encourage reflection, interchange, and support for improving classroom practice in a way that is flexible enough to account for differences among teachers (Harrison 2005; Lee and Wiliam 2005). The materials include 16 modules that form a two-year curriculum for the teacher learning communities (TLCs), participant workbooks for use by these TLCs, and a guidebook for TLC Leaders.

As yet, the *KLT* programme lacks strong effectiveness data and is imperfect in other ways to be discussed in later sections. Its very creation, however, might be viewed as an attempt by Dylan Wiliam and his colleagues to give substantive definition and concrete direction to formative assessment, which some observers felt had the potential to become an empty fad. What provided the potential for making it a fad was a surge in interest among educators, quickly capitalised upon by test publishers (Popham 2006; Shepard 2008), and perhaps by 'process' consultants as well. That surge in interest was, in turn, sparked by very strong claims for effectiveness.

The effectiveness issue

The most widely cited source for these strong claims is almost certainly the pair of papers published by Paul Black and Dylan Wiliam, both then of Kings College, London. 'Inside the Black Box' is a brief position piece that appeared in *Phi Delta Kappan* (Black and Wiliam 1998a) and also as a widely distributed pamphlet (Black and Wiliam 1998b). That article summarises a lengthy, meticulous review, 'Assessment and classroom learning', published the same year in this journal, *Assessment in Education* (Black and Wiliam 1998c).

As noted, one or the other of these articles has been used routinely to undergird claims for the effectiveness of formative assessment. For example, one highly respected testing expert writes, '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, 19). For the reader unfamiliar with the concept of effect size, this claim would be roughly double the average growth US children in the upper primary to lower secondary grades would be expected to make on standardised tests in a school year.[3]

A second, respected expert states, '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 cites the same .4 to .7 effect sizes as claimed above.

In a later article, the same author appears to expand the claim, in terms of both 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, 15). On the same page, the effectiveness claim appears to be made stronger still: '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'. Gains of this magnitude would be in the order of *three* to *six* times average US student growth.[4]

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, and Stiggins 2009, 56).

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 rigour. Meta-analysis was originally conceived as a method for describing the empirical results observed in a research literature, though it has since been extended for inferring underlying population parameters (Hunter and Schmidt 2004, 512). 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'. (See, for example, Glass, McGaw, and Smith 1981, for a detailed classic introduction to the method.) 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 control group or appropriately pooled across the groups).[5]

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 summarisation 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.

In this regard, a major concern with the original Black and Wiliam (1998c) review is that the research covered is too disparate to be summarised 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 behaviour, teacher use of tests, and mastery learning systems. That collection is simply too diverse to be sensibly combined and summarised by a single, mean effect-size statistic (or range of mean statistics).

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 (1998c) state 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 rigorous within their own frameworks and purposes, … the underlying differences between the studies are such that any amalgamations of their results would have little meaning. (53)

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 pamphlet and *Phi Delta Kappan* position paper, Black and Wiliam (1998a, 1998b) 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 mischaracterisation that has essentially become the educational equivalent of urban legend.[6] 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. We discuss here several frequently cited examples to illustrate the nature of the concerns posed.

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 principally on (now quite dated) dissertations conducted by Bloom's students. In a comprehensive literature review that included those same studies, Slavin (1987) 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 1/4 sigma, and in studies with standardised 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 generalisable achievement effects in the more realistic range of 20–50% of [a] standard deviation. (207)

A second commonly referenced source, by Nyquist (2003), is far more recent. The relevance of this source to the school context can be immediately questioned because, although rarely noted in advocates' invocations, it focuses on the college-level population (19). Second, the study is an unpublished master's thesis and, as such, is not generally available (including on the Internet). 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 and rejoinder. Finally, as might be expected for a master's thesis, it has significant limitations.[7]

Two individual studies, by Meisels et al. (2003) and by Rodriguez (2004), also have figured among advocates' evidentiary sources (e.g., Arter 2006, 42; Davies n.d.; Glasson 2008; Kahl 2007; Love 2009, 15; Stiggins 2006). 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) is appropriately modest in his claims. The study's analysis is complicated, incorporating many variables, with no clear interpretation possible regarding a cause-and-effect relationship between formative assessment and student achievement. The study does report achievement effects related to 'classroom self-efficacy' and to 'uncontrollable attributions', constructs which 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 advocates' efficacy claims.

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

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 vs. 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' (278).

The feedback research is sometimes interpreted by advocates to mean that the positive results are attributable to practices consistent with formative assessment (e.g., qualitative characterisations 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) findings, it would appear to be an oversimplification. For example, Kluger and DeNisi note that feedback effects '… are moderated by the nature of the task … [and] the exact task properties that moderate [these] effects are still poorly understood' (275). In a more recent review, Shute (2008) writes, 'Within this large body of feedback research, there are many conflicting findings and no consistent pattern of results' (153). Although she does offer a variety of recommendations based on the research, these recommendations often vary according to *student* characteristics. Shute notes that '… the specific mechanisms relating feedback to learning are still mostly murky, with few (if any) general conclusions' (156).

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 – including a theory of action

and a concrete instantiation – is essential to helping to abstract a class of things to study and make claims about. 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. Unless we understand the mechanisms responsible for change, we won't know if the effects are due to those mechanisms or to irrelevant factors. We also won't be able to predict the conditions, or population groups, for which the formative assessment is likely to work.

The theory of action postulated for *KLT* is particularly instructive because it, perhaps unintentionally, obscures two key mechanisms. The top box in the 'Teacher Outcomes' portion of Figure 1 indicates that 'Teachers elicit evidence of student learning minute-to-minute and day-by-day'. What makes something an 'assessment', however, is not just that evidence is *elicited*. Assessment entails carefully designing situations (or asking questions) so that the elicited evidence can be connected to critical components of domain understanding, an issue to which we will shortly return. Second, presuming that the resulting evidence is relevant, assessment involves making *inferences* from that evidence (Pellegrino, Chudowsky, and Glaser 2001, 42). In this case, those inferences concern what students know and can do, and are used for adapting instruction. That distinction, between making evidence-based inferences and subsequently adapting instruction, is crucial.

The distinction is crucial because a failure in either step can reduce the effectiveness of formative assessment. If the inferences about students resulting from formative assessment are wrong, the basis for adjusting instruction is weakened. Similarly, if the inferences are correct but instruction is adjusted inappropriately, learning is also less likely to occur.

Focusing on these mechanisms suggests that, to be considered effective, formative assessment requires at least two types of argument as part of the theory of action: a *Validity Argument* to support the quality of inferences and instructional adjustments, and an *Efficacy Argument* to support the resulting impact on learning and instruction. Either argument on its own is not enough.

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 and that a certain intervention would be a sensible next step). 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 characterising what a student knows and can do (e.g., with a 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.[8,9]

In contrast to the validity argument, the efficacy argument makes claims about changes in student skill associated with the use of formative assessment. This efficacy claim is not simply that learning will result but, drawing upon the theory of action, that specific mechanisms will cause that result, in particular actions the teacher (or student) takes based on assessment inferences. Thus, logically, the efficacy argument for formative assessment must include the validity argument.[10]

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 (e.g., for

KLT, that teachers attended TLC meetings and spent time sharing and critiquing their formative assessment practices). Second, that backing should include data suggesting that other intermediate outcomes stipulated by the theory of action were achieved (e.g., that teachers actually shared learning intentions, structured opportunities to activate students as instructional resources for one another). Finally, the backing should entail data indicating that students participating in formative assessment changed more in a positive direction on 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).

It should be obvious, then, that data are required to support the theory of action underlying *any* specific approach to formative assessment. It should be equally obvious that every user of formative assessment need not collect such data. To evaluate the theory, data need only be gathered from a reasonably representative subset of those using the approach in question. The goal is to obtain sufficient data from enough contexts to make the validity and efficacy arguments credible, thereby allowing generalised claims for the meaning of formative assessment results and for the impact on learning of using those results instructionally. The standard of rigour being advocated is a scientific one, similar to that required for the effectiveness claims behind any educational intervention.

If we accept that formative assessment programmes do, in fact, require an efficacy argument and an encapsulated validity argument, a related question concerns whether or not those arguments must have a specific substantive focus, the subject of our next section.

The domain dependency issue

This issue concerns whether formative assessment can be maximally effective if theory and development are focused at a domain-independent level. To place the issue in context, we know from cognitive-scientific research that general and specialised knowledge function in close partnership (Perkins and Salomon 1989). By themselves, domain-independent strategies, such as breaking up a complex problem into smaller parts, are broadly useful but weak, serving mostly in the handling of routine problems. Similarly, domain-specific knowledge is powerful but brittle. On its own, such knowledge is effective under very constrained conditions. When the nature of the problem changes such that the bounds of the domain are breached, that knowledge, by itself, is no longer sufficient.

Following this reasoning, to be maximally effective, formative assessment requires the interaction of general principles, strategies, and techniques *with* reasonably deep cognitive-domain understanding. That deep cognitive-domain understanding includes the processes, strategies and knowledge important for proficiency in a domain, the habits of mind that characterise the community of practice in that domain, and the features of tasks that engage those elements. It also includes those specialised aspects of domain knowledge central to helping students learn (Ball, Thames, and Phelps 2008; Shulman 1986).

This claim has at least two implications. The first implication is that a teacher who has weak cognitive-domain understanding is less likely to know what questions to ask of students, what to look for in their performance, what inferences to make from that performance about student knowledge, and what actions to take to adjust instruction. The second implication is that the intellectual tools and instrumentation we give to

teachers may differ significantly from one domain to the next because they ought to be specifically tuned for the domain in question (Hodgen and Marshall 2005).

A possible approach to dealing with the domain dependency issue is to conceptualise and instantiate formative assessment within the context of specific domains. Any such instantiation would include a cognitive-domain model to guide the substance of formative assessment, learning progressions to indicate steps toward mastery on key components of the cognitive-domain model, tasks to provide evidence about student standing with respect to those learning progressions, techniques fit to that substantive area, and a process for teachers to implement that is closely linked to the preceding materials and, therefore, to the domain in question.

In reading, for example, the cognitive-domain model created by O'Reilly and Sheehan (2009) suggests that one key component of proficiency is the ability to use and understand text conventions for various genres – persuasive, literary, informative. For the literary genre, such a convention would be the ability to use and understand plot structure as a comprehension aid. A hypothesised learning progression for that ability would include the following steps: (1) determine the basic idea of plot; (2) identify key plot elements (e.g., climax, resolution); and (3) understand how events related to the plot advance the author's goals. A question linked to the first step might ask students to summarise the plot for a given text, and a domain-specific technique for gathering additional evidence might be to have students complete a graphic organiser calling for the identification of plot elements for text of the teacher's choosing.

This approach implies that formative assessment should be essentially curriculum embedded, a position that Shepard (2006, 2008) has espoused and Shavelson (2008) has illustrated. But how tightly linked formative assessment must be to any given curriculum is unresolved. It may be workable, for instance, to provide formative assessment materials for the key ideas or core understandings in a domain, which should be common across curricula. That would leave teachers to either apply potentially weaker, domain-general strategies to the remaining topics or, working through the teacher learning communities, create their own formative materials, using the provided ones as models.

The measurement issue

A basic definition of educational measurement is that it involves four activities: designing opportunities to gather evidence, collecting evidence, interpreting it, and acting on interpretations. Although programmes that target the development of teachers' assessment literacy cover much of this territory (e.g., Stiggins et al. 2006), the formative assessment literature gives too little attention to that third activity, in particular to the fundamental principles surrounding the connection of evidence – or what we observe – to the interpretations we make of it. This problem was touched upon earlier in the context of the effectiveness issue, when it was noted that formative assessment is not simply the elicitation of evidence but also includes making inferences from that evidence. Because this idea is so foundational, and only just beginning to become integrated into definitions of formative assessment (e.g., Black and Wiliam 2009, 9), we return to it now.

Formative assessment, like all educational measurement, is an *inferential* process because we cannot know with certainty what understanding exists inside a student's head (Pellegrino, Chudowsky, and Glaser 2001, 42; Glasersfeld, as cited in Black and Wiliam 2009, 17–18). We can only make conjectures based on what we observe

from such things as class participation, class work, homework, and test performance. Backing for the validity of our conjectures is stronger to the extent we observe reasonable consistency in student behaviour across multiple sources, occasions, and contexts. Thus, each teacher-student interaction becomes an opportunity for posing and refining our conjectures, or hypotheses, about what a student knows and can do, where he or she needs to improve, and what might be done to achieve that change.

In her 'Classroom Assessment' chapter from the fourth edition of *Educational Measurement*, Shepard (2006) touches on what might be called a 'formative hypothesis' (Bennett and Gitomer 2009). Shepard writes:

> I see a strong connection between … formative assessment practices … and my training as a clinician when I used observations to form a tentative hypothesis, gathered additional information to confirm or revise, and planned an intervention (itself a working hypothesis). (642)

Kane (2006), in the 'Validation' chapter from the same volume, echoes the idea:

> By examining … student work…, the teacher can form hypotheses about the student's competencies and about gaps in … understanding … If a particular set of conjectures … does account for the student's pattern of performance (including mistakes), and no plausible alternative hypothesis does as well, the proposed conjectures can be accepted as a reasonable conclusion about the student. (49)

The centrality of *inference* in formative assessment becomes quite clear when we consider the distinctions among errors, slips, misconceptions, and lack of understanding. An error is what we *observe* students to make – some difference between a desired response and what a student provides. The error we observe may have one of several underlying causes. Among other things, it could be a slip – that is, a careless procedural mistake; or a misconception, some persistent conceptual or procedural confusion (or naive view); or a lack of understanding in the form of a missing bit of conceptual or procedural knowledge, without any persistent misconception. Each of these causes implies a different instructional action, from minimal feedback (for the slip), to reteaching (for the lack of understanding), to the significant investment required to engineer a deeper cognitive shift (for the misconception).[11] The key point, however, is that any attribution of an underlying cause is an inference, a 'formative hypothesis', that can be tested through further assessment. That further assessment might, for example, involve asking for the student's explanation as to why he or she chose to respond in a particular way (thereby making the student a partner in formative assessment); administering more tasks and looking for a pattern of responses consistent with the hypothesis; or relating the error to other examples of the student's performance.

It is worth noting that the generation and testing of hypotheses about student understanding is made stronger to the extent that the teacher has a well developed, cognitive-domain model. Such a model can help direct an iterative cycle, in which the teacher observes behaviour, formulates hypotheses about the causes of incorrect responding, probes further, and revises the initial hypotheses. In addition, if the underlying model is theoretically sound, it can help the teacher discount student responding that may be no more than potentially misleading noise (e.g., slips that have no deep formative meaning).

Formative inferences are not only subject to uncertainty, they are also subject to systematic, irrelevant influences that may be associated with gender, race, ethnicity,

disability, English language proficiency, or other student characteristics. Put simply, a teacher's formative actions may be unintentionally biased. A teacher may more or less efficaciously judge student skill for some, as opposed to other, groups (Bennett et al. 1993), with consequences for how appropriately instruction is modified and learning facilitated. Such an outcome is easily imagined when the teacher's only language is English and the student is an English language learner. For these students, errors in mathematical problem solving may sometimes be rooted in the student's linguistic, rather than conceptual, misunderstanding of the presented problem or lesson (Martiniello 2008), a subtlety the teacher may easily miss.

We can, then, make formative assessment more principled, from a measurement perspective, by recognising that our characterisations of students are inferences and that, by their very nature, inferences are uncertain and also subject to unintentional biases. We can tolerate more uncertainty, and even bias, in our inferences when the consequences of misjudgement are low and the decisions based upon it are reversible. Such conditions are certainly true of formative contexts. That said, the more certain and unbiased we are, the more effectively we can adjust instruction – why spend time trying to correct a misconception when the error was just due to a procedural slip or to a linguistic misunderstanding? Given that fact, we should try our best to decrease uncertainty and bias by considering data from multiple sources, occasions, and contexts; by grounding action in a sound cognitive-domain model, ideally one that accounts for key differences among student groups; and, where possible, by getting input from others as to the meaning of responses from student groups about which we are less knowledgeable.

The professional development issue

Much of the literature on formative assessment conceptualises it as an activity essentially rooted in pedagogical knowledge (e.g., Black and Wiliam 1998c) – i.e., as simply the process of good teaching. I have argued that such a conceptualisation needs also to include reasonably deep cognitive-domain understanding and knowledge of measurement fundamentals. My claim, in essence, is that a subset of these three competencies is unlikely to work.

If this claim is true, how can we best develop teachers' formative assessment practice? A key question in this regard is whether the components can be effectively addressed semi-independently. For example, *KLT* focuses predominantly on the pedagogical-knowledge aspect of the practice. Formative-assessment pedagogical knowledge is connected to domain understanding through the discipline-centred teacher learning communities (see also Harrison 2005). However, deep domain understanding is unlikely to result, if it's not already present, because such understanding is not formally incorporated into the programme. Rather, the development of domain understanding is seen as a 'bonus feature', as opposed to a targeted programme goal (Wiliam and Thompson 2008, 74). Measurement fundamentals, also, are not directly addressed in any systematic way.

The pedagogical-knowledge approach may well be sensible from a practical perspective. Intentionally trying to develop pedagogical knowledge, deep domain understanding, and measurement fundamentals simultaneously may be more than any one professional-development programme can reasonably deliver. At the least, preservice teacher education has a central role to play in developing a firmer foundation upon which in-service programmes can subsequently build.

A related issue is time. Even if we can find a practical way to help teachers build pedagogical skill, deep domain understanding, and a sense of the measurement fundamentals, teachers need significant time. They need time to put that knowledge, skill, and understanding to practice, for example, to learn to use or adapt purposefully constructed, domain-based, formative-assessment materials. Such materials might include items, integrated task sets, projects, diagnostic tests, and observational and interpretive guides. Teachers also need time to reflect upon their experiences with these materials. If we can get teachers to engage in iterative cycles of use, reflection, adaptation, and eventual creation – all firmly rooted in meaningful cognitive-domain models – we may have a potential mechanism for helping teachers better integrate the process and methodology of formative assessment with deep domain understanding.

The system issue

This last issue may be the most challenging of all. The 'system issue' refers to the fact that formative assessment exists within a larger educational context. If that context is to function effectively in educating students, its components must be coherent (Pellegrino, Chudowsky, and Glaser 2001, 255). Gitomer and Duschl (2007) describe two types of coherence, internal and external. Assessment components can be considered *internally coherent* when they are mutually supportive; in other words, formative and summative assessments need to be aligned with one another. Those components must also be *externally coherent* in the sense that formative and summative assessments are consistent with accepted theories of learning, as well as with socially valued learning outcomes. External coherence, of course, also applies to other system components, including pre-service training institutions which must give teachers the fundamental skills they need to support and use assessment effectively. In any event, if these two types of coherence are not present, components of the system will either work against one another or work against larger societal goals.

A common reality in today's education systems is that, for practical reasons, summative tests are relatively short and predominantly take the multiple-choice or short-answer formats. Almost inevitably, those tests will measure a subset of the intended curriculum, omitting important processes, strategies, and knowledge that cannot be assessed efficiently in that fashion (Shepard 2008). Skill integration and strategic coordination, for example, are likely to be given short shrift. Also almost inevitably, classroom instruction and formative assessment will be aligned to that subset and, as a consequence, the potential of formative assessment to engender deeper change will be reduced.

Thus, the effectiveness of formative assessment will be limited by the nature of the larger system in which it is embedded and, particularly, by the content, format, and design of the accountability test (Bennett and Gitomer 2009). Ultimately, we have to change the system, not just the approach we take to formative assessment, if we want to have maximum impact on learning and instruction. Changing the system means remaking our accountability tests and that is a very big challenge indeed.

Conclusion

The term, 'formative assessment', does not yet represent a well-defined set of artefacts or practices. A meaningful definition requires a theory of action and one or more concrete instantiations. When we have those components in place, we have something

useful to implement and to study. The *KLT* Program (ETS 2010) offers such a definition for one category of formative assessment. More work like that is needed to push the field forward.

Second, a more circumspect interpretation of the effectiveness research would be that the general practices associated with formative assessment can, under the right conditions, facilitate learning. However, the benefits may vary widely in kind and size from one specific implementation of formative assessment to the next, and from one subpopulation of students to the next. (As an example, consider the extensive variation in the effectiveness of feedback.) Also, the magnitude of commonly made quantitative claims for the efficacy of formative assessment is suspect, to say the least. 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 advocates cite. Finally, the validity argument, and evidence to support it – both of which should logically be key to any theory of action of formative assessment – are generally absent. Given these facts, as researchers we need to be more responsible in our efficacy claims and, as educators, less immediately accepting of those who push too self-assuredly for quick adoption.

Third, rooting formative assessment in pedagogical skills alone is probably insufficient. Rather, formative assessment would be more profitably conceptualised and instantiated within specific domains. For example, in a special issue of *Applied Measurement in Education*, Shavelson and his colleagues describe embedding formative assessment in a widely used curriculum, *Foundational Approaches in Science Teaching* (Shavelson 2008). ETS' CBAL (Cognitively Based Assessment of, for, and as Learning) initiative, which is building assessments from cross-curricular cognitive-domain models, offers a second example (Bennett 2010).

Fourth, formative assessment entails making inferences about what students know and can do. Therefore, formative assessment is *assessment*, at least in part. This fact implies that relevant measurement principles should figure centrally in its conceptualisation and instantiation. Incorporating measurement principles doesn't mean that validity should be sacrificed for reliability, as some advocates fear, or that inappropriate psychometric concepts, methods, or standards of rigour intended for other assessment purposes should be applied. But it does mean we should incorporate, rather than ignore, the relevant fundamental principles.

Fifth, teachers need substantial knowledge to implement formative assessment effectively in classrooms. It is doubtful that the average teacher has that knowledge, so most teachers will need substantial time and support to develop it. Additionally, teachers will need useful classroom materials that model the integration of pedagogical, domain, and measurement knowledge (e.g., developmentally sequenced tasks that can help them make inferences about what students know with respect to key domain competencies, and about what next to target for instruction).

Finally, we must account for the fact that formative assessment exists in an educational context. Ultimately, we have to rethink assessment from the ground up as a coherent system, in which formative assessment is a critical part, but not the only critical part.

A suitable closing for this paper comes from Shavelson (2008). Referring to his experience creating, implementing, and studying the effects of formative assessment, he writes:

> After five years of work, our euphoria devolved into a reality that formative assessment, like so many other education reforms, has a long way to go before it can be wielded masterfully by a majority of teachers to positive ends. (294)

In other words, 'formative assessment' is both conceptually and practically still a work-in-progress. That fact means we need to be more sensible in our claims about it, as well as in our expectations for it. That fact also means we must continue the hard work needed to realise its considerable promise.

Acknowledgements

I am grateful to Steve Chappuis, Joe Ciofalo, Terry Egan, Dan Eignor, Drew Gitomer, Steve Lazer, Christy Lyon, Yasuyo Sawaki, Cindy Tocci, Caroline Wylie, and two anonymous reviewers for their helpful comments on earlier drafts of this paper or the presentation upon which the paper was based; to Brent Bridgeman, Shelby Haberman, and Don Powers for their critique of selected effectiveness studies; to Dylan Wiliam, Jim Popham and Rick Stiggins for their willingness to consider differing points of view; and to Caroline Gipps for suggesting (however unintentionally) the need for a paper such as this one.

Notes

1. Influential members of the group have included Paul Black, Patricia Broadfoot, Caroline Gipps, Wynne Harlen, Gordon Stobart, and Dylan Wiliam. See http://www.assessment-reform-group.org/ for more information on the Assessment Reform Group.
2. How does formative assessment differ from diagnostic assessment? Wiliam and Thompson (2008, 62) consider an assessment to be diagnostic when it provides information about what is going amiss and formative when it provides guidance about what action to take. They note that not all diagnoses are instructionally actionable. Black (1998, 26) offers a somewhat different view, stating that: '… diagnostic assessment is an expert and detailed enquiry into underlying difficulties, and can lead to a radical re-appraisal of a pupil's needs, whereas formative assessment is more superficial in assessing problems with particular classwork, and can lead to short-term and local changes in the learning work of a pupil'.
3. 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, and Frisbie 2001), and the *Stanford Achievement Test Series Tenth Edition* (Pearson 2004).
4. Stiggins is reported to no longer stand by the claims quoted here (S. Chappuis, April 6, 2009, personal communication). I have included them because they are published ones still frequently taken by others as fact. See Kahl (2007) for an example.
5. Cohen (1988, 25–7) considers effects of .2 to be small, .5 to be medium, and .8 to be large.
6. 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 effects across individual studies would, in fact, be wider than the oft-quoted .4 to .7 range of effects, as each meta-analytic mean itself represents a distribution of study effects. But more fundamentally, the construction of any such range would seem specious according to Black and Wiliam's (1998c) very own critique – i.e., '… the underlying differences between the studies are such that any amalgamations of their results would have little meaning' (53).
7. A partial list of concerns includes confusing association with causation in the interpretation of results, ignoring in the interpretation the finding that results could be explained by (irrelevant) method factors, seemingly computing effect sizes before coding the same studies for the extent of use of formative assessment (introducing the possibility of bias in coding), giving no information on the reliability of the coding, and including many dated studies (57 of the 86 included articles were 30 or more years old) without considering publication date as a moderator variable.

8. The replicability of inferences and adjustments may be challenging to evaluate. It would be easiest to assess in team-teaching situations in which both teachers might be expected to have a shared understanding of their classroom context and students. Outside of team contexts, replicability might be evaluated through video recording of teachers' formative assessment practice; annotation of the recording by those teachers to indicate their inferences, adjustments, and associated rationales; and review of the recordings and annotations by expert teachers for reasonableness.
9. Kane (2006, 23) uses 'interpretive argument' to refer to claims and 'validity argument' to refer to the backing. For simplicity, I've used 'validity argument' to refer to both claims and backing.
10. One could certainly conceptualise the relationship between the validity and efficacy arguments the other way around; that is, with the efficacy argument being part of a broader validity argument, a formulation that would be consistent with Kane's (2006, 53–6) views. Regardless of which argument is considered to be overarching, there is no disagreement on the essential point: both arguments are needed.
11. As suggested, there are other possible underlying causes for student error, some of which may be cognitive and others of which may be affective (e.g., not trying one's hardest to respond). Black and Wiliam (2009, 17) suggest a variety of cognitive causes, including misinterpretation of language, question purpose or context, or the requirements of the task itself. Affective causes may be situational ones related, for instance, to the type of feedback associated with a particular task or teacher, or such causes may be more deeply rooted, as when a student's history of academic failure dampens motivation to respond even when he or she possesses the requisite knowledge. Boekaerts (as cited in Boekaerts and Corno 2005, 202–3) offers a model to explain how students attempt to balance achievement goals and emotional well-being in classroom situations.

Notes on contributor

Randy Elliot Bennett is Norman O. Frederiksen Chair in Assessment Innovation in the Research & Development Division at Educational Testing Service in Princeton, New Jersey. He has conducted research on integrating advances in cognitive science, technology, and measurement to create new approaches to assessment. Since 2007, Bennett has directed an integrated research initiative titled: Cognitively-Based Assessment of, for, and as Learning (CBAL) (http://www.ets.org/research/topics/cbal/initiative). This initiative is attempting to create a model for a balanced system of K–12 assessment that provides accountability information and supports classroom learning. His and his colleagues' work has been described in the George Lucas Educational Foundation publication Edutopia, in Education Week's Teacher Beat online blog, in a *Science* review of innovative approaches to educational assessment, and in the US Department of Education's National Educational Technology Plan 2010.

References

Arter, J. 2006. Making use of data: What educators need to know and be able to do. In *Beyond NCLB: From measuring status to informing improvement*, ed. J. O'Reilly, 39–72. Proceedings of the National Association of Test Directors 2006 symposium. Boone, NC: National Association of Test Directors. http://natd.org/files/uplink/2006proceedings.pdf (accessed February 25, 2009).

Ball, D.L., M.H. Thames, and G. Phelps. 2008. Content knowledge for teaching: What makes it special? *Journal of Teacher Education* 59, no. 5: 389–407.

Bennett, R.E. 2010. Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives* 8, nos. 2–3: 70–91.

Bennett, R.E., and D.H. Gitomer. 2009. Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In *Educational assessment in the 21st century,* ed. C. Wyatt-Smith and J. Cumming, 43–61. New York: Springer.

Bennett, R.E., R.L. Gottesman, D.A. Rock, and F.M. Cerullo. 1993. The influence of behavior and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology* 85, no. 2: 347–56.

Black, P. 1998. *Testing, friend or foe? The theory and practice of assessment and testing.* London: Routledge/Falmer Press.
Black, P., and D. Wiliam. 1998a. Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan* 80, no. 2: 139–48.
Black, P., and D. Wiliam. 1998b. *Inside the black box: Raising standards through classroom assessment.* London, UK: Kings College, London School of Education.
Black, P., and D. Wiliam. 1998c. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice* 5, no. 1: 7–74.
Black, P., and D. Wiliam. 2003. 'In praise of educational research': Formative assessment. *British Educational Research Journal* 29, no. 5: 623–37.
Black, P., and D. Wiliam. 2009. Developing a theory of formative assessment. *Educational Assessment, Evaluation and Accountability* 21, no. 1: 5–31.
Bloom, B.S. 1969. Some theoretical issues relating to educational evaluation. In *Educational evaluation: New roles, new means. The 63rd yearbook of the National Society for the Study of Education, part 2 (Vol. 69)*, ed. R.W. Tyler, 26–50. Chicago, IL: University of Chicago Press.
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, no. 6: 4–16.
Boekaerts, M., and L. Corno. 2005. Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology: An International Review* 54, no. 2: 199–231.
Cech, S.J. 2007. Test industry split over 'formative' assessment. *Edweek* 28, no. 4: 1, 15. http://www.edweek.org/ew/articles/2008/09/17/04formative_ep.h28.html (accessed February 6, 2009).
Chappuis, J., S. Chappuis, and R. Stiggins. 2009. Formative assessment and assessment for learning. In *Meaningful measurement: The role of assessments in improving high school education in the twenty-first century,* ed. L.M. Pinkus, 55–76. Washington, DC: Alliance for Excellent Education. http://www.all4ed.org/files/MeanMeasCh3Chappuis Stiggins.pdf (accessed August 3, 2009).
Cohen, J. 1988. *Statistical power analysis for the behavioral sciences.* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.
Corcoran, T., F.A. Mosher, and A. Rogat. 2009. *Learning progressions in science: An evidence-based approach to reform (RR-63).* New York: Consortium for Policy Research in Education (CPRE), Teachers College, Columbia University.
Davies, A. N.d. Summary of classroom assessment research. Assessment for learning: An online resource for educators. http://annedavies.com/assessment_for_learning_arc.html (accessed February 25, 2009).
ETS (Educational Testing Service). 2009. *Research rationale for the Keeping Learning on Track® program.* Princeton, NJ: Author. http://www.ets.org/Media/Campaign/12652/rsc/pdf/KLT-Resource-Rationale.pdf (accessed December 17, 2010).
ETS (Educational Testing Service). 2010. *About the KLT program.* Princeton, NJ: Author. http://www.ets.org/Media/Campaign/12652/about.html (accessed December 17, 2010).
Furtak, E.M., M.A. Ruiz-Primo, J.T. Shemwell, C.C. Ayala, P.R. Brandon, R.J. Shavelson, and Y. Yin. 2008. On the fidelity of implementing formative assessments and its relation to student learning. *Applied Measurement in Education* 21, no. 4: 360–89.
Gitomer, D.H., and R.A. Duschl. 2007. Establishing multilevel coherence in assessment. In *Evidence and decision making. The 106th yearbook of the National Society for the Study of Education, Part I*, ed. P.A. Moss, 288–320. Chicago, IL: National Society for the Study of Education.
Glass, G.V., B. McGaw, and M.L. Smith. 1981. *Meta-analysis in social research.* Beverly Hills, CA: Sage.
Glasson, T. 2008. Improving student achievement through assessment for learning. *Curriculum Leadership* 6, no. 31. http://www.curriculum.edu.au/leader/improving_student_achievement,25374.html?issueID=11603 (accessed February 25, 2009).
Harcourt Educational Measurement. 2002. *Metropolitan8: Technical manual.* San Antonio, TX: Author.
Harrison, C. 2005. Teachers developing assessment for learning: Mapping teacher change. *Teacher Development* 9, no. 2: 255–63.

Hodgen, J., and B. Marshall. 2005. Assessment for learning in English and mathematics: A comparison. *The Curriculum Journal* 16, no. 2: 153–76.

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

Hunter, J.E., and F.L. Schmidt. 2004. *Methods of meta-analysis: Correcting error and bias in research findings.* 2nd ed. Thousand Oaks, CA: Sage.

Kahl, S. 2007. Formative assessment: An overview. Presentation at the Montana Office of Public Instruction 'Assessment Toolkit' conference, April 23, in Helena, MT. http://opi.mt.gov/PDF/Assessment/conf/Presentations/07MON_FormAssmt.ppt (accessed February 11, 2009).

Kane, M.T. 2006. Validation. In *Educational measurement,* 4th ed., ed. R.L. Brennan, 17–64. Westport, CT: American Council on Education/Praeger.

Kluger, A.N., and A. DeNisi. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin* 119, no. 2: 254–84.

Lee, C., and D. Wiliam. 2005. Studying changes in the practice of two teachers developing assessment for learning. *Teacher Development* 9, no. 2: 265–83.

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

Martiniello, M. 2008. Language and the performance of English-language learners in math word problems. *Harvard Educational Review* 78, no. 2: 333–68.

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

Meisels, S.J., S. Atkins-Burnett, Y. Xue, D.D. Bickel, and S. Son. 2003. Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. *Education Policy Analysis Archives* 11, no. 9. http://epaa.asu.edu/epaa/v11n9/ (accessed February 11, 2009).

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

O'Reilly, T., and K.M. Sheehan. 2009. *Cognitively based assessment of, for and as learning: A framework for assessing reading competency* (RR-09-26). Princeton, NJ: Educational Testing Service.

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

Pearson. 2005. Achieving student progress with scientifically based formative assessment: A white paper from Pearson. http://www.pearsoned.com/RESRPTS_FOR_POSTING/PASeries_RESEARCH/PA1.%20Scientific_Basis_PASeries%206.05.pdf (accessed June 26, 2009).

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

Perkins, D.N., and G. Salomon. 1989. Are cognitive skills context bound? *Educational Researcher* 18, no. 1: 16–25.

Popham, W.J. 2006. Phony formative assessments: Buyer beware! *Educational Leadership* 64, no. 3: 86–7.

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, no. 1: 1–24.

Rohrer, D., and H. Pashler. 2010. Recent research on human learning challenges conventional instructional strategies. *Educational Researcher* 39, no. 5: 406–12.

Sadler, D.R. 1989. Formative assessment and the design of instructional systems. *Instructional Science* 18, no. 2: 119–44.

Scriven, M. 1967. The methodology of evaluation. In *Perspectives of curriculum evaluation,* ed. R.W. Tyler, R.M. Gagne, and M. Scriven, 39–83. Chicago, IL: Rand McNally.

Shavelson, R.J. 2008. Guest editor's introduction. *Applied Measurement in Education* 21, no. 4: 293–4.

Shepard, L.A. 2006. Classroom assessment. In *Educational measurement*, 4th ed., ed. R.L. Brennan, 623–46. Westport, CT: American Council on Education/Praeger.

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

Shulman, L.S. 1986. Those who understand: Knowledge growth in teaching. *Educational Researcher* 15, no. 2: 4–14.

Shute, V.J. 2008. Focus on formative feedback. *Review of Educational Research* 78, no. 1: 153–89.

Slavin, R.E. 1987. Mastery learning reconsidered. *Review of Educational Research* 57, no. 2: 175–213.

Stiggins, R.J. 1999. Assessment, student confidence, and school success. *Phi Delta Kappan* 81, no. 3: 191–8. http://www.pdkintl.org/kappan/k9911sti.htm (accessed February 11, 2009).

Stiggins, R. 2006. Assessment for learning: A key to motivation and achievement. *Edge* 2, no. 2: 3–19. http://www.michigan.gov/documents/mde/Kappan_Edge_Article_188578_7.pdf (accessed February 11, 2009).

Stiggins, R.J., J.A. Arter, J. Chappuis, and S. Chappuis. 2006. *Classroom assessment for learning: Doing it right—using it well.* Princeton, NJ: Educational Testing Service.

Weeden, P., J. Winter, and P. Broadfoot. 2002. *Assessment: What's in it for schools?* London: Routledge/Falmer.

Wiliam, D., and M. Thompson. 2008. Integrating assessment with learning: What will it take to make it work? In *The future of assessment: Shaping teaching and learning,* ed. C.A. Dwyer, 53–82. New York: Erlbaum.

Zwick, R., D. Senturk, J. Wang, and S.C. Loomis. 2001. An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice* 20, no. 2: 15–25.