Cognitively Based Assessment of, for, and as Learning (CBAL): A Preliminary Theory of Action for Summative and Formative Assessment

Randy Elliot Bennett

Educational Testing Service

CBAL (Cognitively Based Assessment of, for, and as Learning) is a research initiative intended to create a model for an innovative K–12 assessment system that documents what students have achieved (of learning); helps identify how to plan instruction (for learning); and is considered by students and teachers to be a worthwhile educational experience in and of itself (as learning). Because CBAL intends to not only measure student achievement but also facilitate it, CBAL, like any similar assessment program, requires a theory of action. This paper describes the notion of theory of action, offers a preliminary version of such a theory for CBAL, and outlines a provisional research program for evaluating that theory.

In his paper introducing this special issue, Michael Feuer calls for a “methodologically rigorous apparatus to assess . . . the expected benefits and costs of test-based accountability programs.” Such an apparatus, of course, presumes that we know in advance what the benefits of any such program are intended to be. This idea is incorporated in the Race to the Top Assessment Program’s Notice of Intent to Apply, which requires state consortia to explicate such intended benefits as part of their proposed system’s “theory of action” (Department of Education, 2010, April 9). The present article describes the notion of theory of action in the context of educational assessment; sketches a preliminary theory for CBAL, an experimental assessment system; and outlines a provisional research plan for evaluating that theory.

THEORY OF ACTION IN EDUCATIONAL ASSESSMENT

Theory of action is a common notion in the program evaluation literature, going back at least to Wholey (1979) and appearing to have come about because program managers were too often unclear about the intended goals of their efforts. The term is closely associated with logic.
model, a graphical or textual depiction of an intervention that explains the cause-effect relationships among inputs, activities, and intended outcomes. Such models are extensively used today in program planning, management, and evaluation throughout the human-services sector as a representation of a program’s theory of action.

The notion of theory of action is not often applied to assessment programs because such programs are not generally intended to cause change in individuals or institutions in the same sense as with an educational or social services intervention. However, in educational accountability testing, as, for example, conducted under No Child Left Behind (NCLB), change is intended, at least at the school level where sanctions for poor performance are applied. As such, it seems appropriate to require a theory of action for such assessment programs, in addition to the more usual scientific evidence in support of instrument technical adequacy. Marion and Perie (2009) offer an example of such a theory for NCLB alternate assessments based on alternate achievement standards.

Modern validity theory appears to subsume the concept of theory of action within the notion of validity argument (Haertel, 1999; Kane, 2006). From this perspective, the consequences that a testing program is intended to produce become part of that argument. (The inclusion of unintended consequences in that argument, however, remains a matter of debate [e.g., Geisinger, 2010].) In modern validity theory and its application, far more attention has been directed at the instrument technical-quality concerns more central to assessment programs than to program effects. One might just as well reverse the relationship and subsume the validity argument within the theory of action (and its associated research program). This alternative formulation arguably gives greater prominence to the effects of the assessment system on individuals and institutions (as well as to the underlying mechanisms behind those effects). Both formulations direct attention to instrument technical quality and to system consequences. The difference between the formulations is in which of those elements is placed in the foreground and in the types of consequences generally included.

With that as context, a theory of action for an assessment system might include the following elements:

- the intended effects of the assessment system
- the components of the assessment system and a logical and coherent rationale for each component, including backing for that rationale in research and theory
- the interpretive claims that will be made from assessment results
- the action mechanisms designed to cause the intended effects
- potential unintended negative effects and what will be done to mitigate them

In the following section, the notion of theory of action is applied to CBAL, followed by a provisional research plan to evaluate that theory.

A PRELIMINARY THEORY OF ACTION FOR CBAL

CBAL, an acronym for Cognitively Based Assessment of, for, and as Learning, is a research initiative funded by ETS (Bennett & Gitomer, 2009). The initiative is intended to create a model for an innovative K–12 assessment system that documents what students have achieved (of learning); facilitates instructional planning (for learning); and is considered by students and teachers to be
a worthwhile educational experience in and of itself (as learning). The model is directed at satisfying the needs of states and state consortia, as well as needs at the district, school, classroom, and individual levels.

Because CBAL is a research initiative, its work has necessarily focused on only some of the subject areas and student populations that an operational assessment system would need to address. The current research focuses on reading, writing, and mathematics for middle-school students being educated to unmodified content standards. Extension to other subjects and grades, as well as to special populations, would need to be added to complete the model.

Figure 1 summarizes the preliminary theory of action for CBAL. The figure shows the CBAL components, action mechanisms, and intended effects. The figure is intended to serve as an advance organizer for the reader, with the understanding that the unfamiliar terms and concepts will be explicated through the discussion below.

![Figure 1](image-url)
Because intended effects should drive system design, we begin our discussion with the “effects” portion of the figure and then move to the CBAL components designed to achieve them. Also discussed (but not represented in the figure) are the interpretive claims, some of which are implicit in the hypothesized action mechanisms, and unintended effects.

Some Intended Effects

As shown in Figure 1, CBAL’s intended effects can be segmented into intermediate effects and ultimate ones to which the intermediate effects are hypothesized to contribute. Among the intermediate intended effects are

- a clearer, deeper, and better-organized understanding on the part of teachers of the content domain in which they teach;
- an increase in teachers’ pedagogical knowledge and assessment skill;
- greater focus in classroom instruction on integrated performances and other activities intended to promote the development of higher-order skills (without neglecting the development of lower-level component skills);
- the routine use of formative assessment practices in the classroom to make appropriate adjustments to instruction; and
- improved student engagement in learning and assessment.

Among the ultimate effects are

- improved student learning with respect to content standards and
- more meaningful information for policy makers regarding the effectiveness of education at the school, district, and state levels, leading to decisions that facilitate learning.

Components of CBAL

CBAL has competency-model, summative-assessment, formative-assessment, and professional-support components (see Figure 1). These 4 components are designed to work synergistically to produce the intended effects. Prototypes that exemplify the components have been developed in collaboration with teachers from the Portland (ME) Public Schools over a 3½-year period.¹

The competency models undergird the other 3 components. These domain-specific models integrate content standards with learning-sciences research and include learning progressions

¹It should be noted that there is no one-to-one mapping between the CBAL components and the notions of “assessment of learning,” “for learning,” and “as learning.” The competency model and professional support components cannot be mapped because they are not assessments but rather complements to assessment. Although summative assessment is frequently equated categorically with “assessment of learning” and formative assessment with “assessment for learning,” the relationships are, arguably, more complex (Bennett, in press). In CBAL, the summative assessments are designed primarily to document what students know and can do, an “of learning” role, and, secondarily, to aid instructional planning and model good learning practice, “for” and “as learning” roles, respectively. CBAL formative assessments are designed primarily as instructional planning aids and as learning experiences but, secondarily, to help teachers form judgments about what students know and can do.
The competency model as a bridge between standards, learning sciences research, and practice.

from that research (where such progressions are available). The rationale for having a competency model is that it acts as a bridge from content standards to test design, curriculum, professional development, and, ultimately, classroom practice. The competency model not only helps specify content but also helps identify research-based instructional- and learning-practice principles. Finally, it conceptually connects summative and formative assessment with one another, and indirectly with curriculum and professional development. Figure 2 shows these relationships.

We have produced CBAL competency models for each of the 3 middle-school domains in which our research is currently being conducted—reading, writing and mathematics. Initial versions of the models can be found in Deane et al. (2008), Graf (2009), and O’Reilly and Sheehan (2009a). The models work in conjunction with content standards to drive domain definition and test design.

Figure 3 shows a highly simplified version of the CBAL writing competency model (Deane, in press). The model emphasizes that writing competency is not synonymous with text-production skills. Rather, the writing competency needed for students to become college ready and career ready requires significant command of, and coordination with, reading and thinking skills. Assessments that attempt to measure writing in isolation are apt to produce an indicator of a de-natured skill that neither gives policy makers the information they need to improve education, nor students and teachers a meaningful target for classroom instruction. The CBAL reading and math competency models similarly incorporate critical thinking skills in ways that current state assessments do not, providing a foundation for more rigorous and conceptually demanding tests.

The CBAL summative assessment component consists of multiple “periodic assessments” distributed across the school year and aggregated together for state and school accountability purposes. The rationale behind this design is fourfold. First, the design can provide a timely status and predictive measure, replacing the need for lower-stakes interim assessment on which students may be less motivated to perform. Second, in the aggregate, the multiple assessments
should allow deeper coverage than a single end-of-year test could afford. Third, because of their frequency, the assessments can keep teacher and student attention more continuously focused on key competencies, as well as on learning principles that CBAL tests purposely model. Finally, the design reduces the impact of a single, end-of-year test because summative decisions are based on multiple pieces of evidence, thereby reducing the effect on students, teachers, and schools of unrepresentative student performance.

Distributing summative assessment across the school year presumes that each periodic assessment will measure some unique competencies and some that are common to the series of measures as a whole. Such a design presumes a common curricular sequence, which all participating schools follow. The granularity of that sequence depends on the number of periodic assessments, though subject area may also play a role (e.g., reading and writing may be less sensitive to instructional order than math). A reasonable number of assessments might be between 2 and 6 per year. A 2-assessment design (i.e., midterm and final), for example, requires specifying only the competencies to be covered for each half-year, leaving the sequence within half-years to local decision makers. Current CBAL designs in reading, writing, and math include 4 periodic assessments and, thereby, may impose greater constraints on curricular order.

The periodic assessments that make up the summative component are standardized and computer-delivered. In CBAL reading and math, the test design for each of these assessments includes one or more extended, scenario-based task sets, plus discrete constructed-response and selected-response items. The reason for this test design is that a mix of question formats should provide more comprehensive coverage of domain-relevant processes, strategies, knowledge, and
habits of mind than a single question format (Messick, 1994). Scenario-based task sets might better target depth-of-processing and some higher-order skills, whereas short elemental items might better provide breadth-of-coverage. In addition, this design, coupled with the aggregation of performance across measures taken at multiple points in time, should ameliorate the generalizability problem so often noted with performance assessments (Linn & Burton, 1994).

For CBAL writing, the current test design contains only a scenario-based task set because the time needed to engage students in a meaningful, extended writing exercise may not be enough to allow incorporation of an additional section of unrelated, discrete items.

Figure 4 shows the introduction to a scenario-based task set from CBAL writing. In this scenario, the student becomes a member of an imaginary team and is called upon to organize, edit, critique, and contribute to the work of other “team members.” Figure 5 gives the 4 major tasks that comprise the set. Each task itself contains one or more related items, with 32 constructed-response and selected-response questions in all. In general, CBAL scenario-based task sets culminate with an exercise that asks the student to respond to the problem posed in the scenario’s introduction. For CBAL writing, the culminating task requires the student to produce a document (or portion of a document) calling for the rhetorical strategies and thinking skills associated with a particular writing purpose and genre, in this case a pamphlet.

Each CBAL periodic assessment is intended to be a worthwhile educational experience in and of itself. The scenario-based task sets, for example, typically require deep engagement with content so that by the time the assessment concludes, the student has learned something nontrivial. Second, each periodic assessment is designed to model one or more aspects of good instructional or learning practice by including tools and representations that proficient performers work with, and by encouraging the habits of mind common to proficient performers in the domain.

FIGURE 4 The introduction to a scenario-based task set for CBAL writing summative assessment.

Note. Copyright © 2010 by Educational Testing Service. Used by permission. Created by Paul Deane and Mary Fowles.
example, CBAL reading assessments include items that model how to analyze complex text by using a graphical organizer. CBAL writing assessments incorporate a variety of planning tools of the type routinely used by practiced performers. Finally, most CBAL assessments present conventional criteria (e.g., for evaluating the quality of a summary, of the information presented on the Internet, of a good persuasive essay) and ask students to apply those criteria as part of assessment tasks. The inclusion of such criteria on the summative assessments (as well as in formative materials) is aimed at encouraging their frequent use to the point that they become a habit of mind. Figure 6 gives an example from a CBAL reading scenario-based task set built around the problem of electronic waste (e-waste).

In contrast to the summative component, which is highly structured, the CBAL formative component is intended to be used by teachers as they see fit. Teachers are encouraged to adapt the formative materials and to create their own materials using the CBAL materials as models. The rationale for the formative component rests upon research suggesting that the general practices associated with formative assessment can, under the right conditions, enhance learning (Bennett, in press; Black & Wiliam, 1998).

Central to this CBAL component is the “formative hypothesis” (Bennett, in press), an inference about what students know and can do that the teacher (or student) confirms or refutes by gathering more information, including through CBAL formative assessment. Formative hypotheses may be initially suggested through various sources of evidence, including the results of CBAL summative assessment, formative assessment, class work, and homework. Some of these same sources may also be used to test and refine the initial hypotheses. Rather than providing a set of scores or a stock diagnosis, the CBAL formative materials are generally intended to give teachers (and students) the evidence needed to build, through posing and testing their formative
FIGURE 6 A CBAL reading summative task requiring students to apply conventional criteria to the evaluation of Internet sources.


hypotheses, a qualitative insight into student understanding which can, in turn, be used to adjust instruction.

Created in collaboration with teachers, CBAL formative assessment includes componential item sets; classroom tasks; extended activities designed to produce evidence about student competency; and teacher guides and interpretive materials. The componential item sets are each targeted at a discrete skill. The sets are intended to help teachers (and students) follow up on poor summative assessment performance, confirm (or refute) areas suggested as needing remediation, and use the results to adjust instruction accordingly. These item sets have a secondary role of helping students build fluency by giving them repeated practice on component skills. In CBAL writing, for example, the competency model links critical thinking skills to the writing genres in which particular skills must be deployed if students are to be effective in that genre. For persuasive writing, those skills include being able to detect which side of an argument an author supports, as well as whether a given piece of evidence supports, refutes, or has neutral impact on a claim. CBAL componential item sets target each of these two skills separately. Figure 7 gives an example set.

In addition to componential item sets, the formative component includes classroom tasks, which preserve the essential elements of the extended, scenario-based task sets found on the periodic summative assessments. The classroom tasks highlight for teachers and students the skill and knowledge constellations required for successful domain performance and their design helps scaffold students toward this successful performance. The tasks can be used to practice putting discrete knowledge-and-skill components together to form more integrated performances; to assess integrated performance and provide evidence to guide adjustments to instruction; and to practice summative test formats.
The third formative element consists of extended activities meant to be carried out over days or weeks. The activities may include projects or other forms of sustained work. These activities are intended to allow for formative assessment of knowledge and skill integration, collaboration, and other competencies that may not be possible to measure through less intensive methods. The activities are also intended to encourage students to take responsibility for a significant self-directed learning experience.

The formative component also contains teacher guides and interpretative materials, including scoring rubrics for constructed-response items and for extended activities. The rationale for including these guides is that teachers need significant domain, pedagogical, and assessment expertise to use formative assessment effectively. By attempting to aid the development of that expertise, the guides provide a concrete link to the last CBAL component, professional support.

CBAL professional support is intended to assist teachers in understanding and implementing the formative and summative assessments, and to help them develop the knowledge and skill needed to do so effectively. This component is implemented through local communities of practice that use the CBAL teacher training guides and interpretive materials, as well as the formative materials, as the basis for their development activities.

**Provisional Interpretive Claims**

To this point, the theory of action has described CBAL’s intended effects and components. This section lists the provisional interpretive claims we would like to be able to make from CBAL summative and formative results. The claims are labeled as “provisional,” and should remain so.
labeled, until such time as sufficient evidence is gathered through the research program to support them.

For summative assessment, some of the provisional interpretive claims are the following:

- Aggregated student performance on periodic assessments represents achievement of content standards.
- Students who perform at a designated level of proficiency are ready to proceed to the next grade’s work.
- Flagged content areas, class groups, and individual students should be followed-up through classroom assessment because they are likely to need attention.
- Achievement, proficiency, and follow-up claims have similar meaning across population groups, including students with disabilities and those who are ELL (taking unmodified tests).
- Districts, schools, and classes that perform poorly on periodic assessments should be followed up because they are likely to need administrative review.

From formative assessment, our provisional interpretive claims include:

- Individual-student and class formative-assessment performance will indicate competencies in need of attention.
- Those competencies can be addressed through adjustments to instruction suggested by the evidence resulting from formative assessment.
- The quality of the inferences suggested and the adjustments made is similar across population groups, including students with disabilities and those who are ELL (being educated to unmodified content standards).

Hypothesized Action Mechanisms

The hypothesized action mechanisms connect CBAL components to the intended effects (see Figure 1); that is, adoption and use of the system components is hypothesized to cause those effects. When CBAL is adopted, we expect that, prior to each periodic summative assessment, teachers will target classroom instruction toward the appropriate CBAL competency model (reading, writing, or math). As part of that process, teachers will communicate to students the learning intentions related to the CBAL competency model and that model’s associated learning progressions. Additionally, we expect that teachers and students will incorporate into their classroom and learning practice the tools and representations characteristic of proficient performers that are found in the formative materials and on the periodic assessments. (As noted earlier, these expectations may presume some degree of common curricular ordering.)

Following each periodic assessment, we expect that teachers will use the results as a starting point for formative follow-up with individual students or groups. This expectation is only reasonable, of course, to the degree that assessment results are available to teachers in a timely manner, targeted competencies are dependent enough on one another to make formative follow-up worthwhile, and teachers have the requisite skills to conduct follow-up effectively.²

²More timely feedback to teachers might be achieved through “phased reporting.” An assessment program might be able, for example, to report initial formative results from a periodic assessment before communicating the summative results that require more time-consuming statistical analysis and demanding quality assurance.
We also hypothesize that summative assessment results will be used by state and local policy makers as a starting point for identifying districts, schools, and classes needing closer examination. That closer examination might include information from a variety of sources: attendance records, teacher observations, school visits, student and parent interviews, and class work. Comprehensive review may lead to providing additional funding, in-service training, or additional staff or to changes in the emphasis of classroom instruction. Ultimately, that review may lead to sanctions of the type included under NCLB.

When CBAL formative materials are used for classroom assessment and instruction, we expect that teachers will reiterate to students the learning intentions related to the CBAL competency model and that model’s associated learning progressions. Teachers (and students) will use CBAL (and other) formative materials to gather information and make inferences about student standing on particular CBAL competency-model components. As part of that process, both teachers and students will score constructed-responses and project work. Teachers and students will use their inferences about student standing to adjust instruction in ways that logically relate to those inferences. Additionally, we expect that teachers and students will use responses to those adjustments, as well as other information, to revise their inferences and readjust instruction. This formative process should be an ongoing one that occurs before as well as after summative assessment. It should also be a process that encourages students to take an active role so that they develop responsibility for their learning and learn to serve as resources for one another (Wiliam, 2007).

Finally, we expect that teachers will regularly meet in discipline-based “communities of practice” to share reflections upon their experiences with CBAL formative-assessment materials, periodic assessment results, and the implications of each for instruction. A significant portion of the community-of-practice interaction would necessarily be centered upon evaluating and discussing examples of student work and the connections of that work to the competency model, associated learning progressions, and the content standards to which the competency model and learning progressions are linked.

Some Potential Unintended Effects

In any assessment program, considerable responsibility rests with the users, who are expected to implement the program as intended and use results in ways consistent with the interpretive claims. As should be clear from the summary depicted in Figure 1, CBAL places even greater responsibility on participants because it also seeks to effect positive change.

By their nature, high-stakes assessment programs are susceptible to unintended effects, or “externalities” (Feuer, 2008), some of which may be associated with the misbehavior of participants. One such effect we might anticipate for CBAL is teaching to the test instead of to the broader construct(s) represented by the relevant CBAL competency model (Koretz & Hamilton, 2006).

A test will always under represent the target construct-domain in at least 2 ways. First, a test is usually a sample of behavior and, in that sense, cannot fully represent the domain to which we wish to generalize. Teaching to that particular sample may increase test performance but not increase performance in the larger domain. Teaching to particular test content—the test items themselves—would consequently be poor instructional practice.
Of more concern to the CBAL theory of action is that a test may be a biased sample of behavior if it privileges particular formats, processes, or strategies to the exclusion of other important formats, processes, or strategies. For purposes of generalizing to the domain, this exclusion may not be problematic if the skills associated with the excluded and included components are highly correlated. From an instructional perspective, however, teaching to the included formats, processes, or strategies may increase test performance but not performance on these other, omitted domain components. Further, over time, even the generalizability of the test may be threatened as students develop more on the practiced components and less on the excluded ones, weakening the association between them. Distributing summative assessment across the school year could potentially exacerbate this situation to the extent that the increased frequency of assessment causes instructional narrowing throughout the year, instead of only toward the end of the year, as might be presumed under a more traditional accountability testing model.

The negative effects of teaching to the test may be somewhat reduced by the explicit identification in CBAL competency models of key domain processes, strategies, and habits of mind. The use of these competency models for test design helps to ensure broader representation of key domain components. Related to this point is the use in CBAL of a more diverse mix of task formats than are typically employed in summative tests. The negative effects of teaching to the test might be reduced further if project work were incorporated into the aggregated summative result. Including project work would further broaden the representation of important processes, strategies, and habits.

A second unintended effect that we might anticipate relates to students from special populations. In creating innovative assessments, we may inadvertently exclude students with disabilities or those who are ELLs because of our choice of task formats, delivery, or other design characteristics. To attempt to mitigate this potential negative effect, we are including, as part of CBAL development, design reviews by experts in assessment procedures used with these groups and cognitive interviews to determine how accessible the tests are for members of these special populations. In addition, we have experimented with incorporating vocabulary links for difficult words (where vocabulary knowledge is not being tested). We may also try out in future pilots the use of alternate representations of the same information (text to speech, described graphics) and alternate questions measuring similar skills at similar difficulty levels, when a class of questions is important but not suitable for some students.

A PROVISIONAL RESEARCH PLAN FOR EVALUATING THE CBAL THEORY OF ACTION

The research plan to evaluate the theory of action for an assessment system might address the following types of questions:

- Is the theory of action logical, coherent, and scientifically defensible?
- Was the assessment system implemented as designed?
- Were the interpretive claims empirically supported?
- Were the intended effects on individuals and institutions achieved, and did the postulated mechanisms appear to cause those effects?
- What important unintended effects appear to have occurred?
The CBAL research initiative is still in the early developmental stages. As a consequence, some of these questions are only beginning to be addressed and others will be addressed only when, and if, the system moves to operational status. What follows is therefore an outline for a research plan to address the 5 questions above, with mention of selected pilot activities to illustrate key elements. However, most pieces of the plan are not yet underway.

Is the Theory of Action Logical, Coherent, and Scientifically Defensible?

Several mechanisms can be used to evaluate the logic, coherence, and defensibility of the theory of action. For example, the theory can be reviewed by a project advisory committee. CBAL has both technical and substantive advisory committees, which meet periodically to review portions of the theory of action. Technical aspects of the CBAL reading, writing, and math assessments have been reviewed by the CBAL Psychometric Technical Advisory Committee, which has helped identify weaknesses in the theory and studies needed to support intended and implicit claims. Substantive aspects for the CBAL writing assessments have been vetted by the CBAL Writing Advisory Committee.

A second potential evaluation mechanism is public presentation, invited critique by independent experts, rejoinder by the proponents, and publication. The intention behind such interchange is to expose the theory broadly to the field so that questionable elements can be either revised or removed, and promising portions pursued and extended. In this spirit, CBAL research staff have made conference presentations focusing on various aspects of the theory of action at the American Educational Research Association, the American Psychological Association, the Cognitive Science Society, the National Council on Measurement in Education, the International Association for Educational Assessment, the National Student Assessment Conference of the Council of Chief State School Officers, and the Society for Text and Discourse, among others. (See Appendix for a selected list of presentations, many of which included reaction by independent discussants.) Also, early formulations of some portions of the theory have been published in research reports and book chapters (e.g., Bennett & Gitomer, 2009; Deane et al., 2008; Graf, 2009; Graf et al., 2009; O’Reilly & Sheehan, 2009a; O’Reilly & Sheehan, 2009b; Sheehan & O’Reilly, in press; Underwood, Zapata-Rivera, & VanWinkle, 2010).

Was the System Implemented as Designed?

System implementation can be evaluated through onsite observation, as well as by interviewing teachers and students and having them respond to questionnaires. Because CBAL is a research activity and not an operational assessment program, the evaluation of system implementation has been limited and informal.

CBAL summative and formative assessment pilot administrations have been conducted in collaboration with the Portland (ME) Public Schools since fall 2007. The initial summative pilots were observed by ETS staff to ensure that the assessments were administered as intended. In Portland, the use of any given set of formative assessment materials can, depending upon the teacher, last between 1 and 10 class sessions. Consequently, questionnaires and telephone interviews have been used primarily, although some observation has been conducted. Additionally several multiweek, formative-assessment, writing and reading pilots have been conducted at ETS.
using local students and teachers over summer vacation periods. Those pilots were observed by ETS CBAL staff. In both these pilots and the ones in Portland, the goal was to understand how teachers and students used the formative materials, and the problems they encountered, so that implementation and the materials themselves could be improved.

Are the Interpretive Claims Empirically Supported?

The collection of initial data to support the claims one intends to make from assessment results should be part of any pre-operational research and development program. The quality of that data will typically be limited, however, by small and unrepresentative samples and by the fact that the instrument designs, themselves, will be evolving. As such, generalizable results will usually not be available until a program is ready to become operational and field-test forms have been administered that possess the essential characteristics of the operational test. Because CBAL is in the early pre-operational stages, this section suggests the types of studies that might be conducted to gather support for the intended and implicit claims.

Summative Assessment Claims

To determine if aggregated student performance on periodic assessments represents achievement of content standards, at least 2 types of studies would seem warranted. One type would be the standard alignment study between test items and content standards. For CBAL, an alignment study against the Common Core State Standards would be particularly timely. A second type of study would entail cognitive interviews to identify whether the processes, strategies, and knowledge students used in item solution were actually the ones test developers intended in their attempts to represent the content standards.

Determining if students who perform at a designated level of proficiency are ready to proceed to the next grade’s work would seem to require a predictive study. That study might look at the relationship between proficiency classifications based on test performance with teacher ratings for the same students made in the subsequent school year.

One might determine in the following manner if flagged content areas, class groups, and individual students should be followed-up through classroom assessment because they are likely to need remediation. For a representative sample of students and of classes, give to a group of expert teachers the summative test results, including total scores, sub scores, and any item responses used in designating content areas, students, or class groups as needing teacher follow-up. Without indicating the designations, ask teachers to independently classify each given student, class group, and content area as needing or not needing formative follow-up. The analysis would include looking at the classification agreement rates among teachers, and between teachers and the summative test, since the meaning of the teacher judgments as a criterion would be stronger to the extent that teachers show consensus among themselves.

The fourth summative assessment claim was that achievement, proficiency, and follow-up claims have similar meaning across population groups, including students with disabilities and those who are ELL (taking unmodified tests). Support for that claim would require in each of the above studies an analysis by population group to ensure that the study results were consistent across those groups (e.g., that aggregated student performance on periodic assessments for boys was as valid an indicator of the achievement of content standards as it was for girls).
The last illustrative interpretive claim we offered was that *districts, schools, and classes that perform poorly on periodic assessments should be followed up because they are likely to need administrative review*. This claim might be evaluated by collecting attendance records, teacher observational results, school-visit reports, student- and parent-interview responses, and class work from a broad sample of classes, schools, and districts, including ones that were high, medium, and low achieving. Experts could then be asked to review each piece of evidence (absent CBAL scores) and rate the entity (i.e., the district, school, class) as to the degree to which it needed administrative review. The interpretive claim would be supported to the extent that the CBAL periodic assessment scores were positively correlated with these expert ratings.

In addition to intended assessment claims, any assessment program will have numerous implicit claims, which are often prerequisite for supporting the intended claims (Haertel, 1999). These claims may be evaluated in part through the examination of empirical data, as well as through independent review of methodology (e.g., through technical advisory committee review).

For CBAL, among the implicit claims are the following:

- **Scores aggregated across periodic assessments can be compared.** One might want, for example, to compare students to one another or to a performance standard, or compare schools. Among other things, support for this claim would require a defensible aggregation and score-linking method. It might also require some assurance that either the CBAL periodic assessments were administered in the same sequence and at approximately the same times, or empirical data suggesting that differences in sequence and timing didn’t materially affect performance.

- **Aggregated scores can be used to measure change.** Policy makers often want to know whether schools are being more effective, a judgment they make in part on differences in test performance between this year’s fifth graders, for example, and last year’s. Among other things, any such comparison implies that the scores from the tests taken at these 2 time points have been linked through some defensible methodology.

  Teachers also often wish to measure student growth within-year and, if scores are reported from periodic assessments taken by students over the course of the year, inferences about growth are likely to be made. Any such inference would require evidence from factor analytic and other studies indicating that the CBAL periodic tests were measuring the same construct on the same scale. Also needed would be some estimate of the confidence that could be placed in difference scores of various magnitudes so that meaningful indications of growth could be highlighted and chance differences deemphasized.

- **Scores from constructed-response tasks, including project work, are generalizable across important facets, such as raters and tasks.** This claim is of particular (but not exclusive) relevance to CBAL writing because each assessment is built around a single extended scenario-based task set. Claims of this type require administering at least 2 such task sets to the same students and having each set scored independently by more than one rater so that the sources of variance can be evaluated.

- **Scores from parallel test forms can be used interchangeably.** Parallel forms of CBAL periodic assessments may be needed for students who are absent from testing, if a security breach has occurred, or if testing is being conducted over an extended time window. Support for the claim would require a defensible score-linking methodology or some other means of achieving comparable scores.
Scores from periodic assessments can be used to identify students at risk of failing to meet proficiency. If there are consequences for schools or students based on achieving a particular level of test performance, teachers likely will use the CBAL periodic assessments to identify students in danger of failing to reach that level. Such use is logical, especially given that the periodic assessments will contribute to the eventual proficiency classification. Support for this use would require a predictive study of the relationship between performance on the periodic assessments and the subsequent proficiency classifications.

Formative Assessment Claims

To support the interpretive claims attached to use of the CBAL formative assessment materials, the following types of studies would need to be done:

- **Individual-student and class formative-assessment performance will indicate competencies in need of attention.** As initial support for this claim, a study might be done in which teachers kept logs describing the inferences they drew from formative results, the rationales behind those inferences, and any other student information that influenced their inferences about competencies in need of attention. The formative assessment results and other student information provided by these teachers (minus their inferences) could then be given to expert teachers for evaluation. The similarity of the inferences drawn by different expert teachers could be compared both among the experts and with the original, generating teacher. Reasonable agreement levels would suggest that experts and the generating teacher see, on the basis of the formative results, the same competencies in need of attention.

- **Those competencies can be addressed through adjustments to instruction suggested by the evidence resulting from formative assessment.** This claim builds on the one above by adding an action implication. With respect to the present claim, we concentrate on the reasonableness of the choice of instructional adjustment, but not its impact, which we will deal with subsequently in the context of intended system effects. Preliminary evidence for the “reasonableness” claim might be derived through an extension to the expert-teacher study described above. After making their inferences from the given formative-assessment results and related student information, the experts would be asked to suggest instructional adjustments. The substantive similarity of these adjustments would then be compared among experts and with the adjustments logged by the generating teachers. Agreement among experts and between the experts and the generating teachers would constitute support for the “reasonableness” claim.

- **The quality of the inferences suggested and the adjustments made is similar across population groups, including students with disabilities and those who are ELL (being educated to unmodified content standards).** Support for this claim would require an analysis, by population group, in each of the above studies to ensure that the results were consistent across those groups.

Were the Intended Effects Achieved?

The intended effects postulated for CBAL are ambitious, to say the least. They essentially promise that the assessment program will not only provide information for education policy
purposes but that, by virtue of its correct implementation, the assessment program itself will have positive effects on individuals and institutions. Evaluating these effects in a way that allows strong causal attributions would be extremely difficult, if not impossible. What should be achievable is to gather qualitative and quantitative data, examine whether those data are at least consistent with the postulated effects, and attempt to discount plausible competing hypotheses for seemingly positive results. Through an ongoing process of gathering evidence, support might either be amassed or the case for the theory of action weakened, suggesting the need for changes to the theory, to the assessment design, or to program implementation.

**Intermediate Effects**

The first intended intermediate effect for CBAL is the development of a clearer, deeper, and better organized understanding on the part of teachers of the content domain in which they teach. At the pilot implementation stage, this effect could be evaluated on a small scale through experimental or quasi-experimental studies that compared teacher lesson plans or teacher concept maps, for those teachers using and those not using the pilot summative and formative assessments. The lesson plans and concept maps would need to be blindly evaluated by domain experts to see if those artifacts indicated a more sophisticated understanding on the part of the experimental group. A complementary approach might include evaluating lesson plans and teacher concept maps before and after the assessment system was introduced on a large scale.

In general, evidence in support of intended effects is made stronger to the extent that the hypothesized action mechanisms appear to be working in the intended manner. In this regard, differences in teacher domain understanding before and after system implementation, or between participating and non-participating teachers, would be more interpretable if those differences were associated, for example, with teacher use of the CBAL competency models, frequent application of the formative assessment materials, and regular participation in CBAL communities of practice that focused on domain-related issues.

A second intended effect for CBAL is an increase in teachers’ pedagogical knowledge and assessment skill. This effect might be evaluated through interviews with teachers, lesson-plan review, and classroom observation in pilot and nonparticipating classes, as well as through larger studies comparing the same group of teachers before and after advent of the assessment system. Of particular interest would be the extent to which teachers could articulate key pedagogical and assessment principles, and describe how they implemented those principles in their classroom practice. Evidence of that implementation from the review of lesson plans and observation of classroom behavior would further strengthen the case.

The third-mentioned effect was of a greater focus in classroom instruction on integrated performances and other activities intended to promote higher-order skills (without neglecting the development of lower-level component skills). This effect also might be evaluated through teacher interview, lesson-plan review, and classroom observation, possibly in conjunction with the studies suggested for the “teacher pedagogical knowledge” effect described earlier. One would hope to see indications of increased classroom emphasis on higher-order skills from, for example, before-to-after introduction of the assessment system.

A fourth effect that also would require studying classroom processes was of the routine use of formative assessment practices in the classroom to make appropriate adjustments to instruction. This effect would likely be a particularly difficult one to detect because “formative assessment
practices” may take multiple forms, including a classroom test with associated student feedback or a real-time interaction between the teacher and the students. As such, video-recording with post hoc review and written commentary by the teacher might be needed to understand what practices the teacher saw as formative and what adjustments to instruction the teacher made as a result. Blind review of the recording and teacher commentary by experts could then be used to rate the quality of teacher-identified formative practices, the reasonableness of the adjustments, and the extent to which CBAL materials were directly employed or otherwise appeared to have influenced teacher behavior. All but the last comparison could be made across teachers who were and were not participating in the assessment system pilots.

A final intermediate effect is of improved student engagement in learning and assessment. Differences in student (and teacher) perceptions about learning and assessment might be compared across groups participating and not taking part in assessment pilots, as well as before and after large-scale implementation. Interviews and questionnaires could be employed. Classrooms could also be observed for differences in engagement indicators.

**Ultimate Effects**

The above effects are postulated to contribute to improved student learning with respect to content standards. This effect is especially hard to evaluate because the obvious outcome measure (i.e., CBAL summative assessment) is stipulated as one of the causes of the effect being evaluated. Learning is hypothesized to occur from the combined impact of preparing for the assessments, taking them, acting on the formative hypotheses they generate, and the resulting changes in teaching and learning practice. As such, this learning effect should be evaluated in multiple ways, including through independent measures. Some possibilities include experimental comparisons of student concept maps and class work for those who did and did not participate in assessment system pilots, as well as comparisons of such artifacts before and after large-scale system implementation. Observational analysis of the relationship between teacher use of formative assessment and student achievement measures might also be done, the hypothesis being that greater teacher use should be associated with higher student achievement. These analyses, as well as those for relevant intermediate effects, should be designed to ensure that impact is evaluated for population groups (e.g., those categorized by gender, race/ethnicity, socioeconomic status, disability, and first language), and not just for students generally.

Last, CBAL is intended to provide more meaningful information for policy makers regarding the effectiveness of education at the school, district, and state levels, leading to decisions that facilitate learning. The information should be more meaningful in significant part because students are more engaged and because teachers have developed greater domain, assessment, and pedagogical expertise. Support for this effect might be garnered through interviews and questionnaires with policy makers, having them blindly compare (if possible) information produced by CBAL with information produced from another assessment program to probe specifically how the information provided by CBAL was (or was not) more meaningful and how it might (or might not) be likely to lead to decisions that enhance learning.

**What Important Unintended Effects Occurred?**

We cited 2 potential unintended effects, teaching to the test instead of to the broader construct(s) represented by the relevant CBAL competency model, and the exclusion of students
with disabilities or who were ELLs. The first unintended effect should be evaluated in conjunction with those studies of intended effects that involve the analysis of teacher interviews, lesson plans, logs, or classroom processes. The second unintended effect might be addressed through the analysis of demographic data on student participation, which should help identify inappropriate exclusion from the regular assessment program.

In general, in the evaluation of intended impact, close attention should be paid to results that suggest the possibility of unintended effects. Any such suggestions could be followed up through questionnaire and interview studies with students, parents, teachers, or administrators to verify the existence of those effects and identify possible solutions for their mitigation.

CONCLUSION

CBAL is a model for an innovative K–12 assessment system that integrates learning-sciences theory with content standards. The model measures achievement through periodic summative assessments that incorporate computer-delivered, scenario-based tasks. Change in students and in teachers is postulated to occur through the use of CBAL competency models, summative assessment, classroom formative assessment, and professional support. In postulating such change, CBAL requires a theory of action and a research program to evaluate that theory.

In this paper, we sketched a preliminary theory of action and a provisional research program for CBAL. To date, we have reported results from a few small-scale trials (e.g., O’Reilly & Sheehan, 2009b) and, as of this writing, are analyzing data from recent larger-scale pilots. Filling out the theory, and meaningfully addressing the many questions needed to evaluate that theory, constitutes a multiyear effort. An extended effort is to be expected because creating an innovative assessment system involves an iterative cycle in which theory is created, instruments are developed and administered, theory and instrumentation are refined, and suitable operational processes are generated only after the basic skeleton of a practical, affordable, fair, and scientifically defensible system emerges. Perhaps one of the great failings of No Child Left Behind (NCLB) was that a detailed theory of action was not formally proposed and evaluated with Education Department support. For NCLB’s successor to have the greatest chance of positive impact, a strong theory of action and a sustained program of rigorous research are essential.

REFERENCES

Department of Education. (2010, April 9). Overview information; Race to the Top Fund Assessment Program; Notice inviting applications for new awards for fiscal year (FY) 2010. Federal Register, 75(68),18171–18185.


APPENDIX

SELECTED CBAL CONFERENCE PRESENTATIONS


Bennett, R. E. (2008, August). Assessment of, for, and as learning: Can we have all three? Paper presented at the International Conference on Teaching and Learning with Technology, Singapore.


