

This article was downloaded by: [Stanford University]

On: 27 September 2008

Access details: Access Details: [subscription number 776101540]

Publisher Routledge

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

APPLIED
MEASUREMENT
IN EDUCATION

Applied Measurement in Education

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title-content=t775653631>

On the Impact of Curriculum-Embedded Formative Assessment on Learning: A Collaboration between Curriculum and Assessment Developers

Richard J. Shavelson ^a; Donald B. Young ^b; Carlos C. Ayala ^{1 c}; Paul R. Brandon ^b; Erin Marie Furtak ^d; Maria Araceli Ruiz-Primo ^e; Miki K. Tomita ^{af}; Yue Yin ^{gh}

^a School of Education, Stanford University, ^b College of Education, University of Hawaii, ^c School of Education, Sonoma State University, ^d Max Planck Institute for Human Development, ^e University of Colorado at Denver and Health Sciences Center, ^f The Curriculum Research and Development Group, University of Hawaii, ^g College of Education, University of Illinois at Chicago, ^h College of Education, University of Hawaii at Manoa,

Online Publication Date: 01 October 2008

To cite this Article Shavelson, Richard J., Young, Donald B., Ayala ¹, Carlos C., Brandon, Paul R., Furtak, Erin Marie, Ruiz-Primo, Maria Araceli, Tomita, Miki K. and Yin, Yue(2008)'On the Impact of Curriculum-Embedded Formative Assessment on Learning: A Collaboration between Curriculum and Assessment Developers',Applied Measurement in Education,21:4,295 — 314

To link to this Article: DOI: 10.1080/08957340802347647

URL: <http://dx.doi.org/10.1080/08957340802347647>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

RESEARCH ARTICLES

On the Impact of Curriculum-Embedded Formative Assessment on Learning: A Collaboration between Curriculum and Assessment Developers

Richard J. Shavelson

*School of Education
Stanford University*

Donald B. Young

*College of Education
University of Hawaii*

Carlos C. Ayala¹

*School of Education
Sonoma State University*

Paul R. Brandon

*College of Education
University of Hawaii*

Erin Marie Furtak

Max Planck Institute for Human Development

Maria Araceli Ruiz-Primo

University of Colorado at Denver and Health Sciences Center

¹Authorship is alphabetical following the main authors.

Correspondence should be addressed to Richard J. Shavelson, School of Education, Stanford University, 485 Lasuen Mall, Stanford, CA 94305-3096. E-mail: richs@stanford.edu

Miki K. Tomita
School of Education
Stanford University
The Curriculum Research and Development Group
University of Hawaii

Yue Yin
College of Education
University of Illinois at Chicago
College of Education
University of Hawaii at Manoa

Assessment of and for learning has occupied center stage in education reform, especially with the advent of the No Child Left Behind Federal legislation. This study examined the formative function of assessment—assessment for learning—recognizing that such assessment needs to be aligned, at least in part, with the summative function of assessment—indexing achievement against standards and progress. Black and Wiliam (1998) suggested that formative assessment might very well improve student learning. Based on these ideas and our own experience with reform science education, we hypothesized that for a small investment of resources we might have a major impact on achievement by embedding formative assessments in a nationally used curriculum. To this end we created a collaboration, described here, between curriculum and assessment developers, created embedded, formative assessments, and studied the impact of science teachers teaching with these materials on middle-school students' motivation, achievement, and conceptual change in a small randomized trial. We also studied the collaboration itself with the intent of informing others who might wish to enter into such collaboration about the potential strengths and challenges experienced. The articles that follow in this special issue report in detail what we did and found out; this article provides a rationale and overview for the study.

INTRODUCTION

In the waves of reform that have swept over education in the past 45 years, assessment has become a major policy lever for improving education through comparisons among schools against standards (assessment's *summative* function). It has also become an instrument for improving classroom teaching and learning (assessment's *formative* function). Indeed, assessment, especially assessment for improving learning, has increasingly been viewed

as an integral part of, no longer separate from, teaching. When the formative and summative functions of assessment are aligned so that the signals about what counts as achievement are consistent to educators, students, parents, and the public, assessment is expected to improve student learning (e.g., Wilson & Bertenthal, 2005).² The research reported in this special issue focuses on the formative function of assessment. Nevertheless, we recognize that what goes into formative assessment for learning needs to be aligned with what policy makers and the public hold teachers and schools accountable for in summative assessment of student achievement. When aligned, not only do we expect enhanced student learning but also opportunities for students to understand what is required of them on standardized achievement tests in high stakes accountability environments by performing on the embedded assessments.

The “big idea” behind this work was that if we could embed assessments in a nationally used curriculum to help guide teaching and learning, and if these assessments had the salutary effect on learning and achievement that research suggested (Black & Wiliam, 1998), then for a relatively small investment (embedding assessments) we might experience a substantial impact on learning and achievement for large numbers of students. Moreover, assuming alignment (at least to some degree) between curriculum assessments and standardized assessments, students might transfer their embedded and end-of-unit test performance skills to standardized testing situations.

Practical Influence

We did not, however, immediately arrive at this big idea. Our work grew out of two lines of influence, one practical and one scholarly. On the practical side, we had worked extensively with teachers and curriculum developers. We realized that with the recognition of assessment as integral to education reform, the stage was set for “a romance (collaboration) between curriculum reform and assessment reform” (Shavelson, 1995, p. 58).

In working with science teachers, especially with teachers in an exemplary inquiry-science elementary school (Shavelson 1995), a central question emerged: “What do we intend students to learn in this unit?” Moreover, we found that teachers needed end-of-unit assessments aligned with inquiry science goals to give them a sense of what students should learn. And they also needed mini-assessments that were coordinated with the end-of-unit assessment and

²Indeed, in some states this alignment is happening, perhaps somewhat less optimally in desperation rather than planned, as summative test-like items are provided to teachers for their use in class in response to the pressures of the NCLB federal legislation.

embedded in between lessons (Shavelson, 1995, p. 62). For instance, students in one exemplary classroom had performed poorly on an embedded “Electric Mysteries” assessment (Shavelson, Baxter, & Pine, 1991). The teacher was beside himself—how could this be? It turned out that he had selectively perceived positive performance as he taught. But when standing back and observing students’ performance on the embedded assessment, not jumping into the teachable moment when students took a wrong turn in the investigation, his saw what students did not know or were not able to do with electric circuits, and that led him to change his teaching.

In other experiences we found that the goal of a particular inquiry science unit was not necessarily clear to teachers, let alone the sequence of sub-goals needed to arrive at the end-of-unit goal. As we noted (Shavelson, 1995, p. 64): “Anyone who observes the most progressive, constructivist science teaching will find a classroom filled with activity. . . . But the astute observer will find that . . . the purpose of the lesson is not necessarily obvious to her.” Indeed, we often found that teachers proceeded through a unit only upon reaching the end to find out what and why they were teaching the activities they did. We reasoned, then, that embedded assessments could be used to *signal* a unit’s goal structure and give direction to teachers.

Enter the assessment developer. “It turns out that not only are performance assessments and constructivist curriculum reform simpatico in the classroom, assessment development can reshape the curriculum itself, by clarifying goals, by identifying inconsistencies in or between lessons, or by identifying extraneous lessons” (Shavelson, 1995, p. 64). Hence, there was firm ground for establishing a romance between curriculum developers and assessment developers. Each brought something unique to the table. Moreover, recognizing that in high stakes accountability environments, students needed to encounter, embedded in their curriculum, the kinds of tasks and reasoning required on statewide achievement assessments, we viewed embedded assessments as such a vehicle. The aim, then, was to align such assessments with inquiry science goals and with the types of test items that tapped these goals. In this way, formative assessment could serve summative needs as well.

Scholarly Influence

The second line of influence was scholarly. This line of work addressed two questions beyond the question, “Where do we want to go?” They are: “Where are we? How do we get where we want to go?” The answer seemed to come in a review of literature published by Black and Wiliam (1998) on the effects of what they called “formative assessment,” on student learning. They reported that formative assessment—assessment that provided *immediate feedback* to students on *how* to improve their learning—produced a large positive effect on students’

learning. They also noted that this kind of feedback rarely occurred in classrooms, that studies of teachers' formative assessment practices were needed, and if classroom formative assessments were effective, that the potential for improving students' learning was substantial.

THE STUDY

If our practical experience and Black and Wiliam were on target, embedding formative assessments in a science curriculum should lead to improved teaching and student learning. To test out this hypothesis, we began a collaboration between curriculum developers in the Curriculum Research & Development Group (CRDG) at the University of Hawaii and assessment developers in the Stanford Education Assessment Laboratory at Stanford University (SEAL). Our goals were twofold: (1) to test out our hypothesis that embedding formative assessments within a science curriculum—CRDG's *Foundational Approaches in Science Teaching* (Pottenger & Young, 1992)—would improve teaching and learning; and (2) to study and evaluate the assessment development process that emerged out of this collaborative relationship.

Before describing the program of research more fully, we note several unique features of the work. First, through a series of iterative studies we refined the embedded assessments. This curriculum-and-assessment development work culminated in a final study that tested the effects of embedded assessments on teaching and students' learning in a small randomized field trial. Second, we went beyond the usual definitions of science achievement as largely acquisition of declarative and procedural knowledge and evaluated the claim that formative assessment promotes *conceptual change*. We conjectured that formative assessment would do so directly and possibly indirectly through enhancing student motivation and/or achievement. Third, we studied the collaboration itself—a “study within a study”—trying to understand its ups and downs with the intent to informing future such attempts. We followed Cronbach and Associates' (1980, p. 214) idea that methods of evaluation (in our case the collaborative methods used to develop and study formative assessment) would improve faster if we (evaluators or researchers) could provide a retrospective perspective on the study choices. And finally, as this special issue documents, we examined formative assessment in greater depth than has been reported in previous empirical work on the topic. Suffice it to say here that the comprehensive analysis of curriculum, the backward mapping of formative assessments onto the scientific investigations, the link between a conception of science achievement and the embedded assessments, and the integration of formative assessment ideas, motivation, achievement, and conceptual change, taken together, are not found, to our knowledge, elsewhere.

FORMATIVE ASSESSMENT

Formative assessment takes place on a continuous basis. It is conducted by the teacher with the intent of informing teacher and students as to the gap between what students know and can do and what they are expected to know and be able to do with immediate, informative feedback (Shavelson, 2006). Classroom formative assessment ranges on a continuum from informal to formal. Where a particular formative assessment practice falls on the continuum depends on the amount of planning involved, its formality, the nature and quality of the data sought, and the nature of the feedback given to students by the teacher. We describe three anchor points on the continuum: (a) “on-the-fly,” (b) planned-for-interaction, and (c) formal and embedded in curriculum (Figure 1; for details, see Furtak & Ruiz-Primo, 2007; Shavelson, 2006; Shavelson et al., 2008). By embedding assessments in the *Foundational Approaches in Science Teaching* (FAST) program’s unit on buoyancy, we created a form of formal formative assessment (e.g., Black & Wiliam, 2004a, 2004b).

On-the-Fly Formative Assessment

On-the-fly formative assessment arises when a “teachable moment” unexpectedly occurs, for example, when a teacher circulating and listening to the conversation among students in small groups overhears a student say that, as a consequence of her or his experiment, “density is a property of the plastic block and it doesn’t matter what the mass or volume is because the density stays the same for that kind of plastic.” The teacher recognizes the student’s grasp of density and challenges the student with other materials to see if she or he and her or his group-mates can generalize the density idea.

Planned-for-Interaction Formative Assessment

Planned-for-interaction formative assessment is deliberate. A teacher plans for and crafts ways to find the gap between what students know and what they need to know. For example, while developing a lesson plan, a teacher may prepare a set of “central questions” that get at the heart of the learning goals for that day’s



FIGURE 1 Variation in formative-assessment practices.

lesson. These questions may be general (“Why do things sink and float?”) or more specific (“What is the relationship between mass and volume in floating objects?”). At the right moment during class, the teacher poses these questions, and through a discussion the teacher can learn what students know, what evidence they have to back up their knowledge, and what different ideas need to be discussed. This contrasts with typical classroom recitation where teachers use simple questions to “keep the show going.”

Embedded-in-the-Curriculum Formative Assessment

Embedded-in-the-curriculum formative assessment comes “ready-to-use”; teachers or curriculum developers place formal assessments ahead of time in the ongoing curriculum to create goal-directed “teachable moments.” These assessments are embedded at junctures or “joints” in a unit where an important sub-goal should have been reached before students go on to the next lesson. Embedded assessments inform the teacher about what students currently know, and what they still need to learn (i.e., “the gap”) so that teachers can provide timely feedback.

Formal embedded assessments used in our study were developed by a team of curriculum developers, assessment developers, scientists, and teachers. They were consistent with curriculum developers’ (CRDG’s) intention for the curriculum and instructional goals, with content expert’s understanding (scientists), with psychometric principles (SEAL), and with practice. Consequently, formal embedded assessments provide thoughtful, curriculum-aligned, and valid ways of determining what students know, rather than leaving the burden of planning and assessing on the teacher alone.

SOME QUESTIONS ADDRESSED

The studies reported in this issue address a number of questions, some of which were intentional—for example, What is the impact of formative assessment on students’ conceptions of sinking and floating?—and some that arose in the process of carrying out the study—for example, If assessments are to be embedded in curricular material and used during teaching, where should they be placed?

To be a bit more explicit, the following sampling of questions arose during the course of the study and are addressed in the articles that follow: (1) What critical issues need to be considered in deciding where to embed formal formative assessments? (2) What characteristics should assessments tasks have to be effective *and practical* as *formative* assessments (see Furtak & Ruiz-Primo, 2007)? (3) How might teaching tools—such as the learning progression students move through in reaching a scientifically justifiable explanation for sinking and floating—be designed to assist teachers in interpreting students’ conceptions as they

progress through the unit embedded with formative assessments? (4) How important is it for embedded assessment to meet high psychometric standards and what tradeoffs are involved? (5) Where do teachers have difficulty in carrying out formative assessment practices (eliciting students' conceptions? getting students to use evidence to justify their conceptual claims?) and how can training address these difficulties?

CONCEPTUAL FRAMEWORK FOR DEVELOPING EMBEDDED FORMATIVE ASSESSMENTS

Embedded assessments are intended to focus teaching and learning on the goals of the curriculum and provide feedback to students as to how to close the gap in their knowledge between what they know and what they need to know (e.g., Black & Wiliam, 2004a, 2004b). What then do we want students to know? The answer to this question is important and needs to be consistent across lessons otherwise the assessments will be haphazard and potentially misleading.

With this logic, we built embedded assessments following SEAL's conceptual framework for (science) achievement. The framework proved heuristic in generating assessments—the kinds of test-like tasks overlapped but went beyond what is typically found in state science assessments—and presaged the 2009 NAEP science assessment (National Assessment Governing Board, 2006). The framework also provided some unexpected pitfalls.

We conceived of science achievement as involving cognition, emotion, and motivation (e.g., Shavelson et al., 2002) but for this study, focused directly on cognition. Nevertheless, we also examined the impact of formative assessment on motivation and emotion. Our working definition of science achievement (Li, Ruiz-Primo, & Shavelson, 2006; Shavelson & Ruiz-Primo, 1999; see also National Assessment Governing Board, 2006) involved four types of knowing and reasoning in a subject matter (Figure 2). One type of such knowledge is “knowing that”—*declarative* (factual, conceptual) knowledge. For example, knowing that force equals mass times acceleration and being able to reason with this knowledge. Achievement also involves “knowing how” to do something—*procedural* (step-by-step or condition-action) knowledge and reasoning with this knowledge. For example, procedural knowledge involves knowing how to get the mass of an object or how to carry out and reason through a comparative investigation by manipulating the variable of interest and controlling others. Achievement also importantly involves “knowing why”—*schematic* (“mental model”) knowledge. Such knowledge builds on and connects declarative and procedural knowledge; it is used to reason about, predict, and explain things in nature. For example, schematic knowledge is involved in explaining why some things sink in water and others float. Finally, achievement involves “knowing when and where to apply

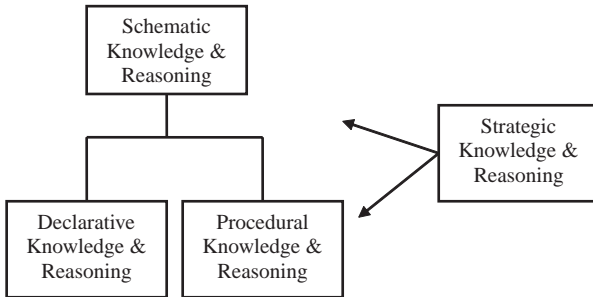


FIGURE 2 Sketch of the knowledge/reasoning types and their relationships.

knowledge,” and to check if the application of this knowledge is reasonable—*strategic* knowledge. For example, experts know when to apply Newton’s first law given a force and motion problem whereas novices are attracted to the surface features of the problem (Chi, Feltovich, & Glaser, 1981).

Although we distinguish these types of knowledge and the corresponding reasoning that goes with them, in reality the separation is not so simple. To reason with schematic knowledge for instance, one typically needs some declarative and procedural knowledge. For example, to reason about what would earth’s climate be if the planet sat perpendicular in its orbit around the sun, all three types of knowledge would be needed. Moreover, to a greater or lesser extent, strategic knowledge is involved whenever a student draws on another type of knowledge in addition to being able to reason strategically, especially in novel situations.

To make this abstract discussion concrete, consider the following as an embedded assessment. The assessment is embedded in a transition or “joint” in the sequence of lessons at the point where students learn to reason about densities. Students have learned about and experimented with the ideas that all else being equal, the more massive an object the greater the sinking in water; the greater the volume, the less sinking; the density of an object is its mass per unit volume; and things more dense than water ($\text{g/cm}^3 > 1$) sink in water and things less dense than water float ($\text{g/cm}^3 < 1$).

At this point, the teacher shows students that a bar of soap sinks in room-temperature water. She or he next cuts the bar into two pieces— $\frac{1}{4}$ and $\frac{3}{4}$. She or he holds the larger piece above the beaker of water and asks students to predict whether it will sink or float and justify their answers. She or he then places the larger piece in the beaker of water and students observe that it sinks. She or he then takes the smaller piece and does the same. The students (some in great surprise) observe the smaller piece to sink. (We call this a predict-observe-explain or POE assessment.) Students are then challenged to explain what they saw.

Their justifications and explanations reveal their schematic knowledge (knowing and reasoning why) and become the focus of classroom discussion with the goal of closing the gap in “mental models” for explaining what they saw, based on empirical data.

FAST UNIT ON BUOYANCY AND JOINTS FOR EMBEDDING ASSESSMENTS

We used the first 12 investigations of a FAST physical science sequence that focused on matter and buoyancy. The sequence develops the concept of buoyancy as the ratio of the object’s density (e.g., plastic cube) to the medium’s density (e.g., water). In developing this notion, the program followed the historical sequence of concept’s development moving from a mass- or volume-only explanation (all else equal) to a combination of mass and volume to the concept of density, to buoyancy as the ratio of two densities.

We identified four natural “joints” in this investigative sequence where assessments might be embedded. At the end of the first four investigations, students should have acquired the notion that the greater the mass, the greater the depth of sinking, all else being equal. Here is where the first embedded assessment was inserted. Beginning with the fifth investigation students then take up the effect of varying volume on sinking and floating, holding mass constant. There is a natural joint between investigations 7 and 8; a short assessment was embedded in the unit to assess students’ understanding of the effect of volume on sinking and floating. By investigation 10, the notion of the object’s density is developed and once again there is a natural joint for embedding an assessment. The soap POE would fit here. And after investigation 12, students are expected to understand that sinking and floating depend on the relative densities of the object and the medium supporting it. At this point, one or more assessments of buoyancy would be embedded before the unit moves onto different objects and media (e.g., gasses).

We developed two types of assessments to be embedded at joints in the buoyancy sequence. One type of assessment focused on procedural and schematic knowledge employing graphing, Predict-Observe-Explain tasks, and constructed responses to the question, “Why do things sink and float?” The second type of embedded assessment—concept mapping (Ruiz-Primo & Shavelson, 1996)³—focused on the structure of students’ declarative knowledge.

³Concept maps are networks where the nodes are key concept terms (e.g., mass, volume, density) and students are asked to connect these terms with arrows showing the direction of relationship between a pair of terms and label the arrow to say how the two terms go together.

DEVELOPMENT AND EVALUATION OF EMBEDDED ASSESSMENTS

SEAL and CRDG jointly developed and evaluated the assessments. An Assessment Development Team (for details, see Ayala et al., this issue)—comprised of teachers, a scientist, science educators, and assessment developers—was responsible for providing a blueprint for the embedded assessments, following our framework (Figure 2). The SEAL staff translated the blueprint into embedded assessments and CRDG staff reviewed them and suggested revisions. The initial blueprint and the draft assessments underwent repeated pilot testing. In the major pilot test, three teachers were briefly trained both in what embedded assessments are and how to use them. At the end of a six-month tryout, the assessment blueprint and assessments were completely revised, as was teacher training in the use of formative assessments (see Ayala et al. and Brandon et al., this issue).

IMPACT OF FORMATIVE ASSESSMENT ON LEARNING, MOTIVATION, AND CONCEPTUAL CHANGE

We evaluated the impact of the final version of the embedded assessments, called “Reflective Lessons,” on teaching and student outcomes in a small randomized trial (Figure 3; see Yin et al., this issue; Furtak et al., this issue).

Participants

Twelve experienced FAST teachers, identified by CRDG as expert in teaching FAST and drawn from across the United States, participated in the experiment; they varied considerably in backgrounds (Table 1). All but one teacher in the study taught more than one section of the FAST program. Although we examined the results of the experiment on all teachers’ classes, in this special issue we concentrate on “focus” classes—classes in which we videotaped each and every lesson in order to study the implementation of the formative assessment

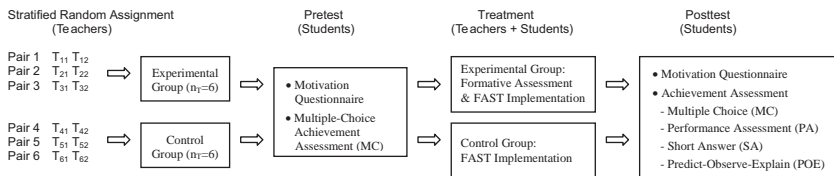


FIGURE 3 Schematic of the small-scale experimental trial.

TABLE 1
Information about the Participant Teachers and Their Classes (from Yin, 2005, Table 3.2, p. 45)

Teacher	Experimental					Control					
	Teaching Experience ^a	Degree	Student Grade Level	Number of Classes ^e	Average Class Size	Teacher	Teaching Experience	Degree	Student Grade Level	Number of Classes	Average Class Size
Aden	Total ^a : 2	BS	7	FAST: 5	29	Rachel	Total: 12	BS	7	FAST: 3	28
	Science ^b : 2	MA		Non-FAST: 1			Science: 12	MS		Non-FAST: 0	
	FAST ^c : 2						FAST: 12				
Andy	FAST I ^d : 2					Lenny	FAST I: 48				
	Total: 5	BA	7	FAST: 2	30		Total: 6	BS	7	FAST: 5	22
	Science: 5			Non-FAST: 3			Science: 3			Non-FAST: 0	
Becca	FAST: 2					Ellen	FAST: 3				
	FAST I: 8						FAST I: 3				
	Total: 18	BE	7	FAST: 5	21		Total: 5	BA	7	FAST: 6	27
	Science: 17			Non-FAST: 0			Science: 5	Minor S		Non-FAST: 0	
	FAST: 12						FAST: 5				
	FAST I: 12						FAST I: 5				

Carol	Total: 23 Science: 10 FAST: 1 FAST I: 1	BA ME	6	FAST: 1 Non-FAST: 5	26	Sofia	Total: 28 Science: 15 FAST: 4 FAST I: 4	BA MA	6	FAST: 3 Non-FAST: 2	27
Diana	Total: 3 Science: 1 FAST: 1 FAST I: 1	BA	6	FAST: 2 Non-FAST: 2	23	Ben	Total: 15 Science: 15 FAST: 11 FAST I: 9	BS MA	7	FAST: 5 Non-FAST: 1	26
Robert	Total: 14 Science: 14 FAST: 7 FAST I: 0	BS	7	FAST: 1 Non-FAST: 4	26	Serena	Total: 22 Science: 6 FAST: 2 FAST I: 2	BS MA	6	FAST: 4 Non-FAST: 0	21

Note. ^aTotal: the total years the teacher has been teaching. ^bScience: the years the teacher has been teaching science. ^cFAST: the years the teacher has been teaching FAST. ^dFAST I: the times the teacher has been teaching FAST I. In some schools FAST I is taught more than once a year. ^eNumber of classes indicates teachers' teaching load. Some teachers also taught non-FAST classes besides FAST classes.

“treatment.” The results reported here are the same as those we obtained with all classes (see Yin, 2005).

Experimental Design

Teacher selection and assignment to group

Teachers were matched pairwise with data provided by CRDG on school district characteristics including school size, percent free lunch, percent proficient on state examination, and ethnic mix (see Yin, 2005, Table 3.1, p. 45). They were then randomly divided into an experimental (6 teachers) and a control group (6 teachers). The experimental group taught the formative-assessment embellished FAST program while the control group taught the regular program (Figure 3).

Teacher training

Both groups received three days of common training that included: (a) study orientation, (b) exchange of how they approached FAST physical science investigations and the specially developed exercises they used to teach about buoyancy; (d) how to use reporting tools (e.g., logs), and (e) how to set up and the use video cameras in their classrooms.

The experimental group teachers were additionally trained in the use of formative assessments (see Ayala et al., this issue, for the substance of the training). For each assessment suite (called “Reflective Lessons”) embedded at a critical joint, the following cycle of training was provided. Teachers first participated as students as project staff administered a Reflective Lesson. They then discussed the Reflective Lesson among themselves and staff, noting the procedural skills needed as well as the role of eliciting students’ conceptions and using those conceptions to build an empirically justifiable knowledge claim. Next, they worked in small groups and taught one another with the Reflective Lessons and received feedback from peers and staff. Then they taught a small group of students studying buoyancy in CRDG’s summer school program, receiving feedback from peers and staff, *as well as from students!* Finally, they reflected on how to improve their administration of and teaching with Reflective Lessons.

Testing

The achievement/conceptual change assessment included questions focusing on declarative, procedural, and schematic knowledge (drawing more or less on strategic knowledge). Example achievement/conceptual change items are included in the Appendix. Motivation and constructed response items are presented in individual papers as appropriate.

THE STUDY WITHIN THE STUDY: THE UPS AND DOWNS OF THE COLLABORATION

In addition to studying our hypothesis that embedding formative assessments in the FAST program would improve students' science motivation and achievement, we set out to learn from the romance between curriculum developers (CRDG) and assessment developers (SEAL; see Brandon et al., this issue). Our goal here was to learn from our experience so as to guide us and others who similarly collaborate on developing formative assessments to embed in a curriculum.

The collaboration spanned three major phases: (1) project design, refinement and assessment team selection; (2) assessment development; and (3) assessment field-test in a randomized experiment. We focused on issues of effective collaboration between the two teams, providing team members' and observers' reflections about the strengths and weaknesses of the collaboration and, when appropriate, suggestions for improving future collaborations.

To be sure we encountered problems in our collaborative development of embedded formative assessments and when training teachers how to use them. For example, the collaboration across 2,400 miles resulted in communication gaps between CRDG and SEAL staff. The groups came to realize that they needed to meet in person frequently, despite cost and time. We also note strengths. For example, CRDG and SEAL staff members were respectful of each other, were task-oriented, and "left their egos at the door."

OVERVIEW OF THE SPECIAL ISSUE

The remainder of this special issue contains four articles. They report on the formative assessment development and teacher training processes, the evaluation of formative assessment's impact on student motivation and learning and possible interpretations of these findings, the fidelity with which the formative-assessment treatment was implemented and their possible interpretations, and on the lessons learned.

The second article (Ayala et al.), "From Formative Assessment to Reflective Lessons to Preparing Teachers to Use Reflective Lessons," describes the process of embedded assessment development, tying the embedded and end-of-unit assessments to the FAST program and a learning trajectory based on FAST and research on cognitive development. We found developing embedded assessments took a great deal of care, that teachers' preconceptions about assessment influenced how they used embedded assessments, and that even in-depth training did not necessarily overcome teachers' preconceptions about assessment and inquiry science teaching.

The third article (Yin et al.), “On the Measurement and Impact of Formative Assessment on Students’ Learning and Motivation,” evaluates the instruments used in the study and reports the results of the randomized experiment. The findings, highlighting the importance of teachers in the formative-assessment equation, showed large variability in teachers’ practices, regardless of treatment condition, which in turn impacted student outcomes.

The fourth article (Furtak et al.), “On the Fidelity of Implementing of Formative Embedded Assessments and Its Relation to Student Learning,” links the findings of large teacher effects on student learning together with teachers’ classroom assessment practices. Of particular importance is the contrast in inquiry teaching practices, regardless of treatment condition, and the teaching strategies that distinguished more and less effective formative assessment practices.

The fifth article (Brandon et al), “On Lessons Learned and Future Collaborations,” draws together the CRDG-SEAL collaborative experiences with the intent of informing and improving future such collaborations. The article concludes with recommendations for future collaborations.

CONCLUDING COMMENTS

This project examined the impact of embedded formative assessments in an inquiry-science curriculum on students’ achievement, conceptual change, and motivation. It tested conjectures based on practice and prior research. If successful, the potential impact would be significant. That is, if by simply embedding formative assessments in curricula student achievement and motivation could be substantially enhanced, we might very well have a cost-efficient way of improving science (and other) education outcomes. Although the prospect is dizzying, the reality that we report here is sobering. Teachers have a huge impact on the efficacy of any education reform, and we can add formative assessment to the list. The gap between what Black and Wiliam and we envisioned and the reality is quite significant. That said, the evidence tentatively suggests that when teachers employ formative assessment practices as intended in this study, student outcomes may be enhanced.

We are heartened by the findings about the collaboration between curriculum developers (CRDG) and assessment developers (SEAL). Both parties learned a great deal from one another. We now know how to do a better job of collaborating and these lessons might very well inform other such collaborations. As for the CRDG-SEAL romance, the collaboration continues. Both groups came to appreciate the importance of teacher engagement for any education innovation no matter how transparent it seems. We are not alone and

once again recognize the importance teacher development plays in education reform.

Finally, we want to point out that what we have developed in the context of formative assessment—the assessment framework linked to assessment items—speaks directly to large-scale, summative assessment. Through this and other research, we are able to produce reliable and valid assessments of students' science achievement that tap into our knowledge framework. This development would serve summative assessment purposes well. And in building formative and summative assessments from the same framework, it just might be possible to align the two assessment functions; the 2009 National Assessment of Educational Progress Assessment Framework and Test and Items Specifications provide an existence proof (http://www.nagb.org/frameworks/naep_science_framework_2009.doc and http://www.nagb.org/frameworks/naep_science_specs_2009.doc).

ACKNOWLEDGMENTS

This research was supported, in part, by a grant from the National Science Foundation (NSF/Award ESI-0095520) and, in part, by the National Center for Research on Evaluation, Standards, and Testing (CRESST/Award 0070 G CC908-A-10).

REFERENCES

- Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5(1), 7–73.
- Black, P., & Wiliam, D. (2004a). Classroom assessment is not (necessarily) formative assessment (and vice-versa). In M. Wilson (Ed.), *Towards a coherence between classroom assessment and accountability* (pp. 183–188). Chicago: University of Chicago Press.
- Black, P., & Wiliam, D. (2004b). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 20–50). Chicago: University of Chicago Press.
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.
- Cronbach, L. J. & Associates (1980). *Toward reform of program evaluation*. San Francisco: Jossey-Bass.
- Furtak, E. M., & Ruiz-Primo, M. A. (2007). *Studying the effectiveness of four types of formative assessment prompts in providing information about students' understanding in writing and discussions*. Paper presented at the American Educational Research Association Annual Meeting, Chicago, IL.
- Li, M., Ruiz-Primo, M. A., & Shavelson, R. J. (2006). Towards a science achievement framework: The case of TIMSS 1999. In S. Howie & T. Plomp (Eds.), *Contexts of learning mathematics and science: Lessons learned from TIMSS* (pp. 291–311). London: Routledge.
- National Assessment Governing Board (2006). *Science assessment and item specifications for the 2009 National Assessment of Education Progress* (pre-publication ed.). Washington, DC: Author.

- Pottenger, F. M. I., & Young, D. B. (1992). *The local environment: FAST 1, Foundational approaches in science teaching*. Honolulu: University of Hawaii Curriculum Research & Development group.
- Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching*, 33(6), 569–600.
- Shavelson, R. J. (2006). On the integration of formative assessment in teaching and learning: Implications for new pathways in teacher education. In F. Oser, F. Achtenhagen, & U. Renold (Eds.), *Competence-oriented teacher training: Old research demands and new pathways* (pp. 63–78). Utrecht, The Netherlands: Sense Publishers.
- Shavelson, R. J. (1995). On the romance of science curriculum and assessment reform in the United States. In D. K. Sharpes & A.-L. Leino (Eds.), *The dynamic concept of curriculum: Invited papers to honour the memory of Paul Hellgren* (pp. 57–76). (Research Bulletin 90). Finland: University of Helsinki, Department of Education.
- Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, 4(4), 347–362, (Special Issue, R. Stiggins & B. Plake, Guest Editors.)
- Shavelson, R. J., Roeser, R. W., Kupermintz, H., Lau, S., Ayala, C., Haydel, A., Schultz, S., Quihuis, G., & Gallagher, L. (2002). Richard E. Snow's remaking of the concept of aptitude and multidimensional test validity: Introduction to the special issue. *Educational Assessment*, 8(2), 77–100.
- Shavelson, R. J., & Ruiz-Primo, M. A. (1999). On the psychometrics of assessing science understanding. In J. J. Mintzes, J. H. Wamhersee, & J. D. Novak (Eds.) *Assessing science understanding: A human constructivist view* (pp. 303–341). New York: Academic Press.
- Shavelson, R. J., Yin, Y., Furtak, E. M., Ruiz-Primo, M. A., Ayala, C. C., Young, D. B., Tomita, M. K., Brandon, P. R., & Pottenger, F. (2008). On the role and impact of formative assessment on science inquiry teaching and learning. In J. E. Coffey, R. Douglas, & C. Stearns (Eds.), *Assessing science learning: Perspectives from research and practice* (pp. 21–36). Washington, DC: NSTA Press.
- Wilson, M. (Ed.) (2004). *Towards a coherence between classroom assessment and accountability*. Chicago: University of Chicago Press.
- Wilson, M. R., & Bertenthal, M. W. (Eds.) (2005). *Systems for state science assessment*. Washington, DC: National Academies Press.
- Yin, Y. (2005). *The influence of formative assessment on student motivation, achievement, and conceptual change*. Unpublished doctoral dissertation. Stanford, CA: Stanford University.

APPENDIX

Example Multiple-Choice Achievement Test Items

Declarative Knowledge and Reasoning:

11. Density equals

- A. buoyancy divided by mass.
- B. buoyancy divided by volume.
- C. volume divided by mass.
- D. mass divided by volume.

Procedural Knowledge and Reasoning:

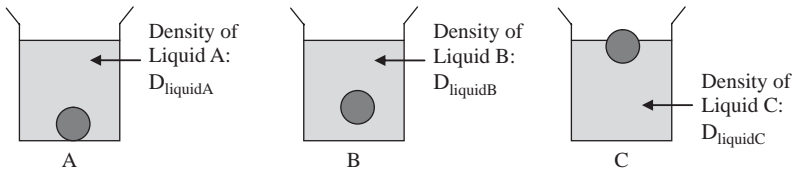
24. Which object listed in the table has the greatest density?

<i>Object</i>	<i>Mass of Object</i>	<i>Volume of Object</i>
W	11.0 grams	24 cubic centimeters
X	11.0 grams	12 cubic centimeters
Y	5.5 grams	4 cubic centimeters
Z	5.5 grams	11 cubic centimeters

- A. W.
- B. X.
- C. Y.
- D. Z.

Schematic Knowledge and Reasoning:

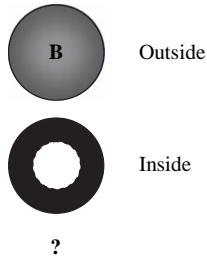
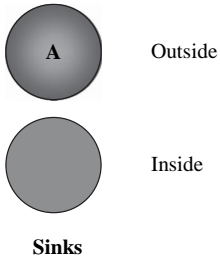
23. Melody has three balls of the same material. The balls have the same mass and volume. She put the balls into three different liquids: A, B, and C. See picture below. What is the relationship between the densities of the three liquids?



- A. $D_{\text{liquidA}} > D_{\text{liquidB}} > D_{\text{liquidC}}$.
- B. $D_{\text{liquidA}} < D_{\text{liquidB}} < D_{\text{liquidC}}$.
- C. $D_{\text{liquidA}} = D_{\text{liquidB}} = D_{\text{liquidC}}$.
- D. $D_{\text{liquidA}} = D_{\text{liquidB}} < D_{\text{liquidC}}$.

Conceptual Change (Schematic Knowledge/Reasoning):

40. Ball A and ball B have the **SAME** mass and volume. Ball A is solid; Ball B is hollow in the center (see the pictures below). Ball A sinks in water. When placed in water, ball B will _____



- A. sink.
- B. float.
- C. subsurface float.
- D. not sure.