The Nature and Impact of Teachers' Formative Assessment Practices

CSE Technical Report 703

Joan L. Herman, UCLA/CRESST
Ellen Osmundson, UCLA/CRESST
Carlos Ayala, CSU, Sonoma
Stephen Schneider, WEST-ED
Mike Timms, WEST-ED
Center for Assessment and Evaluation of Student Learning
CAESL

December 2006

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education.

The National Science Foundation provided support for this research under a grant to WEST-ED for the Center for Assessment and Evaluation of Student Learning (CAESL). However views expressed in the paper do not necessarily represent the views of the Foundation.

THE NATURE AND IMPACT OF TEACHERS' FORMATIVE ASSESSMENT PRACTICES[1]

Joan L. Herman, UCLA/CRESST
Ellen Osmundson, UCLA/CRESST
Carlos Ayala, CSU, Sonoma
Stephen Schneider, WEST-ED
Mike Timms, WEST-ED[2]

Center for Assessment and Evaluation of Student Learning
CAESL

Abstract

Theory and research suggest the critical role that formative assessment can play in student learning. The use of assessment in guiding instruction has long been advocated: Through the assessment of students' needs and the monitoring of student progress, learning sequences can be appropriately designed, instruction adjusted during the course of learning, and programs refined to be more effective in promoting student learning goals. Moving toward more modern pedagogical conceptions, assessment moves from an information source on which to base action to part and parcel of the teaching and learning process. The following study provides food for thought about the research methods needed to study teachers' assessment practices and the complexity of assessing their effects on student learning. On the one hand, our study suggests that effective formative assessment is a highly interactive endeavor, involving the orchestration of multiple dimensions of practice, and demands sophisticated qualitative methods for study. On the other, detecting and understanding learning effects in small samples, even with the availability of comparison groups, poses difficulties to say the least.

Long-standing theory and research suggest the critical role that formative assessment can play in student learning. With roots in Ralph Tyler's curriculum rationale (1949), B.F. Skinner's behaviorism and programmed instruction (1953, 1960),

[1] Paper Prepared as part of Symposium Building Science Assessment Systems That Serve Accountability and Student Learning: The CAESL Model for the annual meeting of the American Education Research Association, Montreal, Canada, April 2005

[2] The authors would like to thank Stephen Zuniga and Sam Nagashima, graduate students at CRESST, UCLA for their help with data analysis.

Robert Glaser's seminal work in criterion referenced instruction and testing (Glaser, 1963), and Benjamin Bloom's concept of Mastery Learning (Bloom, 1968), the use of assessment in guiding instruction has long been advocated: Through the assessment of students' needs and the monitoring of student progress, learning sequences can be appropriately designed, instruction adjusted during the course of learning, and programs refined to be more effective in promoting student learning goals. Moving toward more recent pedagogical theory, Sadler (1989) adds the important cognitive and social functions that assessment can provide in teaching and learning and the significant role that feedback from assessment play in enabling teachers and students to understand their learning goals, to compare the actual level of their performance to the desired level, and to engage in effective actions to reduce the gap.

In modern pedagogical conceptions, in fact, assessment moves from an information source on which to base action to part and parcel of the teaching and learning process. That is, contemporary cognitive psychology recognizes that knowledge is always actively constructed by learners and a situative perspective reminds us that knowing is a verb before it is a noun (NRC, 2001a, 2001b). What is acquired through schooling is a set of capabilities for meaningful participation in activity structures; all knowing has a social component. And socio-cultural perspectives remind us as well of the political, social, and motivational functions of assessment (Gipps, 1999). Assessment itself provides opportunities for students to display their thinking and to be engaged with feedback that can help students to extend, refine, and deepen their understandings and reach more sophisticated levels of expertise. For example, interim assessments or quizzes during the course of instruction or questioning during class discussions can serve to elicit students' thinking, feedback can be used to encourage students to confront their misconceptions, and the process itself can be instrumental in helping students move to higher levels of understanding (Gitomer & Duschl, 1995).

Formative assessment thus serves multiple functions in instruction and learning, and the rationale for its benefits on learning is multifaceted. Recent research, in fact, documents its strong effects. Black and William's landmark meta-analysis of 250 studies addressing various aspects of formative assessment found effect sizes ranging between .4 and .7, leading the researchers to conclude that formative assessment should be considered prime among available interventions for improving student learning (Black & Wiliam, 1998). Their review also shows that formative assessment may be particularly effective as a strategy for improving the learning of low-ability students.

Yet even as research shows the rich potential of formative assessment, so too it suggests the limits of current practice. Available teaching materials lack the types of systematic and sensitive assessments that teachers and students need to both spark and make visible students' thinking and to discern the details of student progress to inform subsequent action. Moreover, teachers and schools have limited background and capacity to engage in assessment (Heritage & Yeagley, 2005; Herman & Gribbons, 2001; Plake & Impara, 1997; Shepard, 2001; Stiggens, 2002). As Black and Wiliam well note, assessments can only become formative when information from them is *used* to adapt teaching and learning for the benefit of student learning.

But what does good *use* of assessment mean in classroom practice? How can and do teachers routinely use classroom assessment to support their learning goals? How can and do teachers orchestrate, synthesize and act upon the informal assessments they may derive "on the fly" (Bell & Cowie, 2001) during classroom interactions with the more formal assessment offered by student work, unit tests, and quizzes? What is the role of assessment in effective instructional practice? We believe these questions have been understudied and would benefit from additional exploration, to which the study below is a response. It uses multiple methods to examine whether and how teachers *use* quality formative assessments, the impact of such practice on student learning, and the factors that influence its efficacy. In the process, our research team seeks an operational definition of "quality assessment practice" and to derive implications for teacher capacity building. The study is part of a larger study by the Center for the Assessment and Evaluation of Student Learning (CAESL) aimed at demonstrating the feasibility and value of integrated assessment systems, where large scale and formative assessment are synchronized on common goals and where assessments are grounded in modern theories of cognition, reflect developmental learning perspectives, and provide useful and action-oriented data on student learning trajectories. In the sections below, we first lay out an initial framework for conceptualizing quality practice, and then describe the methods of our study, its results, and implications.

A Model of Quality Assessment Practice

Figure 1 below summarizes the conceptual model underlying the study. In essence, the CAESL tetrahedron makes a number of assertions. First, it asserts that sound formative assessment must be based on both quality assessment *tools* and quality *use* of information from such tools. We believe this is an important distinction, in that much of the literature addressing teachers' assessment practices addresses one or the

other of these components. For example, the highly regarded "Knowing What Students Know" (KWSK; NRC, 2001b) brought together decades of research in cognition, measurement, and psychometrics to make eloquent arguments about the role of assessment in learning, the importance of teachers' assessments, and the need to create learning-based assessments that integrate available knowledge about cognition and learning with assessment development, measurement theory, and psychometrics. While the scholars participating in KWSK conceptualized a powerful design model for doing so, their charge does not extend to consider whether and how such assessments can be used by teachers to improve their students' learning. At the other end of the continuum, Inside the Black Box (Black & Wiliam, 1999) and Bell and Cowie's case studies of formative assessment (2001), deal with teachers' use of assessment to promote student learning, but largely neglect issues of the quality of the assessments used. While they do consider validity and the match between the curriculum and assessment goals, admittedly a crucial important concern in formative assessment quality (and one that is highlighted in Figure 1), there are many other important dimensions to assure validity. Yet without quality assessment that is valid for its intended uses, data and feedback that are derived from assessments may provide faulty information and result in faulty inferences and inappropriate action.

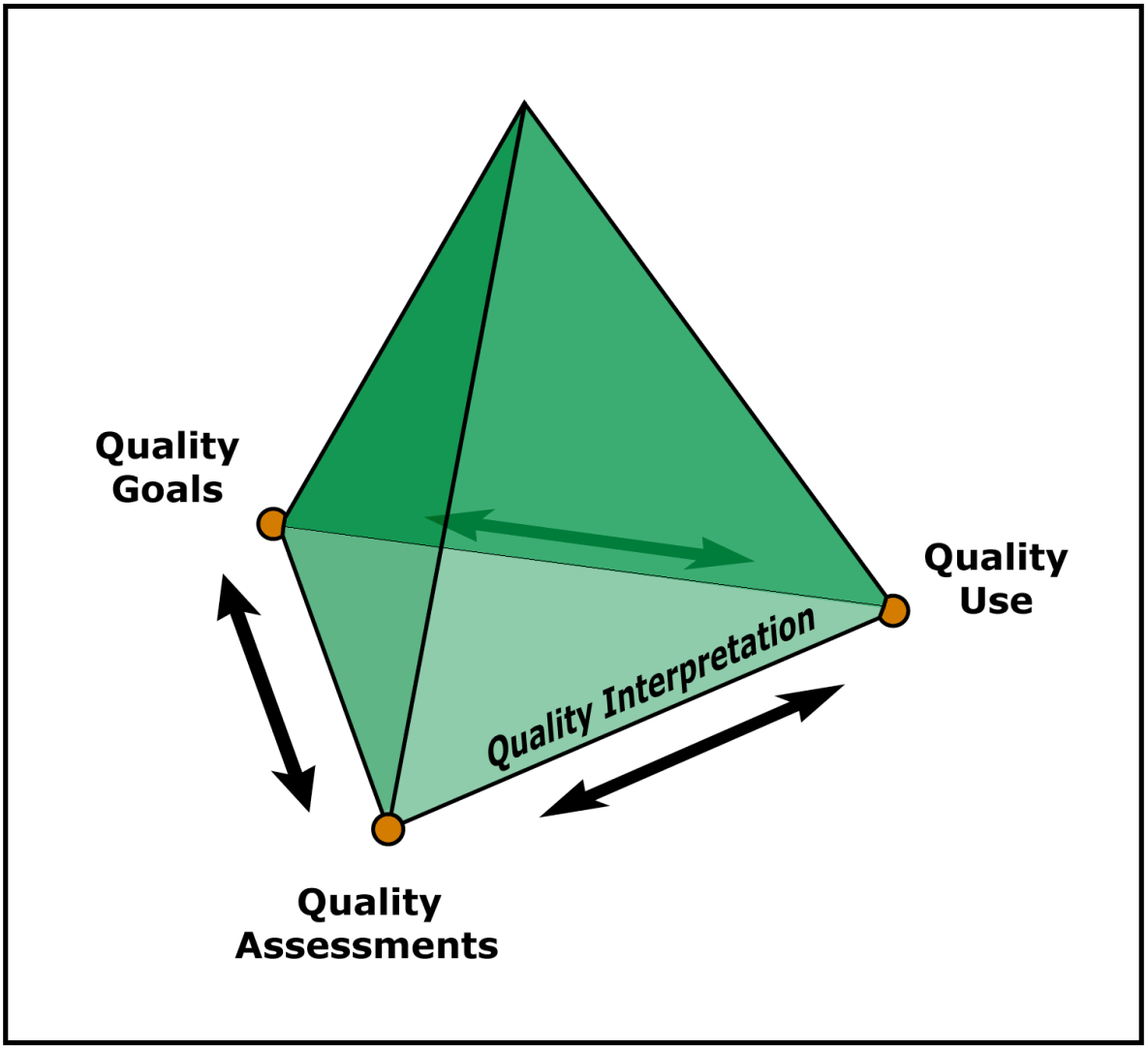


Figure 1. CAESL quality classroom assessment framework.

As noted above, our conceptual model makes clear the primacy of learning goals in quality assessment practice and their link to effective pedagogy. Learning goals are the starting, ending, and recycling points in the selection and implementation of quality assessment tools, in interpretation and analysis of student work, and in the use of results to provide informative feedback and take action that will further students' progress, e.g., by probing and responding to individual students' understanding, those

of subgroups and/or those of the class as a whole; by using results to modify curriculum and instruction for future classes.

Our views of the systemic nature of quality assessment practices also are worth noting. We see quality assessment as an integrated system of information that provides appropriate detail for gauging and responding to student progress on desired learning goals throughout the instructional process. With specified goals and pedagogical pathways for achieving them in the forefront, teachers can be constantly evaluating whether and how students are making expected progress on the way and making adjustments as needed. This means, for example, that learning activity goals on a day-by-day (or on a continual basis) are coordinated with specified unit and/or long-term goals, and that assessment is purposively designed and used—whether it be a classroom discussion question, a probe during individual work, a quiz, or a test—to detect student progress. Mark Wilson and colleagues have coined the term "progress variables" to denote the hierarchies of learning against which students' progress can be monitored; these same progress variables can be used to coordinate large scale and classroom assessment (Wilson & Sloane, 2000). In any event, with such a integrated system, teachers must coordinate and orchestrate between the various levels—that is relating today's data with yesterday's, to larger unit goals, to other available data on student learning—and move back and forth between pedagogical plans and assessments of student progress for both short and associated longer term learning goals.

This conceptual model defining the nature of quality assessment practices underlies the central issues addressed by our study and the instrumentation we used to assess them. However, while the quality of assessment tools is a critical component of our model, the study controlled for it by studying assessment practice in the context of a curriculum with known, quality assessments. As noted earlier, our primary research purposes focused on describing the quality of teachers assessment practices, looking for linkages between such practice and student learning, and deriving implications for professional development and future research.

Methods

We describe below the curriculum unit and its embedded assessments that provided the context for our study. We then follow with a description of our sample, our data sources, and analysis strategies. Our data sources included multiple measures of teachers' knowledge, instructional practices, and student learning.

The FAST Curriculum Unit on Sinking and Floating

We selected as a context for our study the Foundational Approaches in Science Teaching (FAST) for middle-school science, developed by the University of Hawaii Curriculum Research and Development Group (Pottenger & Young, 1992). The program is based on a constructivist philosophy of learning, is aligned with National Science Standards (Rogg & Kahle, 1997), and uses carefully sequenced, student-conducted investigations to develop students' learning. Students often work in small groups to share ideas, observations, and think about results; are actively involved in observation, summary, and drawing conclusions; and encouraged to link their learning to prior experiences. Previous studies have supported the efficacy of FAST (Pauls, Young, & Lapitkova, 1999; Tamir & Yamamoto, 1977; Young, 1993), and FAST has been designated as an exemplary program by the U.S. Department of Education's (2001) Expert Panel on Mathematics and Science Education and by the National Staff Development Council (Killion, 1999).

The study focuses on the first 12 investigations of the introductory Physical Science Strand of FAST1, The Local Environment, one of three FAST texts (see Table 1). The investigations engaged students in a variety of science skills—observation, graphing, summarizing results, providing explanations—and addressed concepts of mass, volume, density, and relative density and their relationship to buoyancy—or our short hand, "Why Things Sink and Float." The unit required 10-12 weeks to implement.

Table 1.

First Twelve FAST 1 Investigations of Student Tasks and Learning Goals

(Adapted from Ruiz-Primo& Furtak, 2004)

Lesson	Investigation	Student Tasks	Learning Goals
1	Liquids And Vials	Observing vials of different liquids sinking and floating in different liquids (a buoyancy anomaly)	Make scientific observations and test predictions
2	Sinking Straws	Adding BBs to a straw and measuring the depth of sinking	Predict the number of BBs needed to sink a straw to particular depth
3	Graphing Sinking Straws Data	Graphing number of BBs versus depth of sinking	Represent BB data in line graphs; more BBs more sinking
4	Mass And The Sinking Straws	Finding the relationship between total mass and depth of sinking of straws	Conclude that more mass more sinking
5	Sinking Cartons	Measuring the depth of sinking of different sizes and of equal mass	Discover the relationship between the amount of ballast, the carton size and the depth of sinking
6	Volume And Sinking Cartons	Finding the submerged and total volume of cartons	Calculate the displaced volume of different cartons
7	Floating And Sinking Objects	Finding the mass and volume of different objects	Graph mass vs. displace volume of floating and sinking objects
8	Introduction To Cartesian Divers	Experimenting with Cartesian Divers	Discover how a Cartesian Diver works
9	Density And The Cartesian Diver	Investigating the relationship between the diver's mass and volume at different sinking and floating positions	Find the density of Cartesian divers of different masses and volumes
10	Density of Objects	Finding the relationship between total volume and sinking and floating; discover density	Find the density of floating and sinking objects; density graph.
11	Density of Liquids	Determining the density of liquids	Discover that different liquids have different densities
12	Buoyancy of Liquids	Experiment with different objects in liquids of different densities	Understand relative density

This unit was particularly chosen for study because of the formative assessments that had been specially developed and embedded in it, called "Reflective Lessons"

(Shavelson, SEAL & CRDG, 2005, p. 6). As described below, the Reflective Lessons were based on Shavelson's theory of achievement testing and had been subjected to considerable validity research. In contrast to the vast majority of available science curriculum, we could be confident about the quality of the unit's formative assessments. They were not an afterthought, but had been specially developed to illuminate different types of student thinking and to provide teachers with feedback on students' learning and potential next steps for teaching and learning.

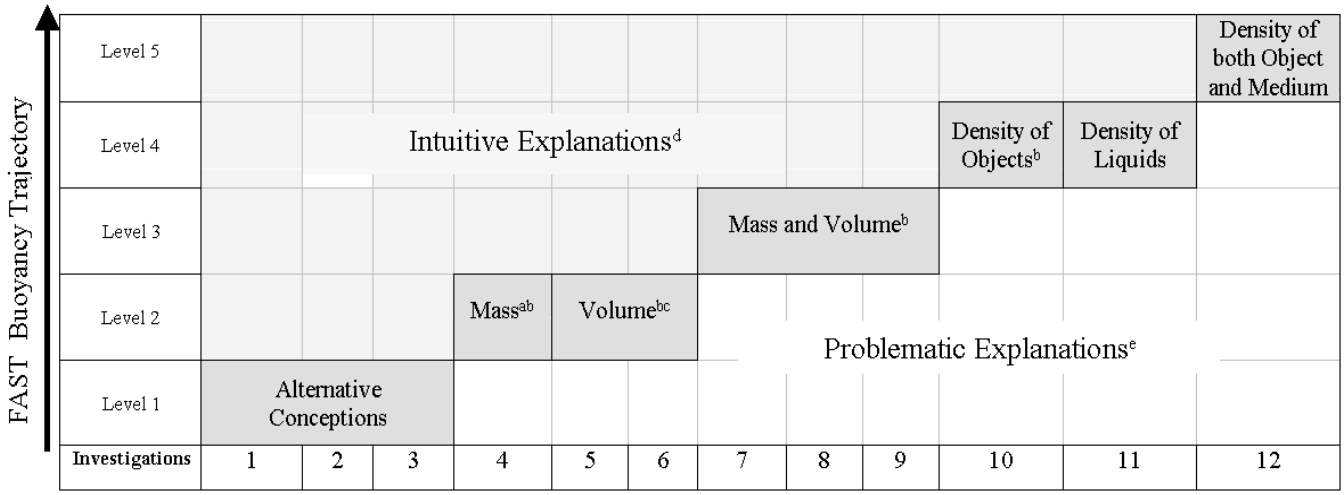
The FAST/CAESL Formative Assessments: Reflective Lessons

Termed "reflective lessons" to avoid expectations for grading and to communicate that they were opportunities for teachers and students to reflect on their learning progress, the formative assessments addressed key juncture points in the curriculum, when students were expected to transition from one sub-goal of the unit to the next and where both teachers and students might benefit from feedback about whether students were ready to move on. These points corresponded to the points at which students might be expected to understand progressively the role of mass, volume, both mass and volume, and relative density in buoyancy.

The reflective lessons were intended to elicit and make public student conceptions, encourage communication and argumentation, monitor and extend student conceptions, and reflect on student understandings. (See Ayala, 2005 for details of rationale and development.) The lessons were of two types. Each Type I reflective lesson suite asked students to (1) interpret and evaluate a graph, (2) predict-observe-explain an event related to sinking and floating, (3) answer a short question, and (4) predict-observe an event related to sinking and floating. These were embedded after Investigations 4, 7 and, 10, and in turn were called RL4, RL7 and RL 10. Type II reflective lessons used Concept Maps to investigate students' progress in their conceptual understanding of key concepts and materials included across the twelve investigations—e.g., the concepts of mass, volume, density, displaced water, floating, sinking, materials of wood, straws, etc. The concept mapping tasks asked students to show the relationship between the various terms as a window on student conceptions. These concept-mapping tasks were inserted after PS6 and PS11 and were respectively called RL6 and RL11.

<u>Progress variables and scoring rubrics.</u> The reflective lessons were scored using rubrics reflecting each of two progress variables underlying the unit. The first progress variable reflected FAST's implicit developmental model for fostering student

understanding of buoyancy, as shown in Figure 2. Students were scored at one of five levels, depending on their explanations for buoyancy. (See Ayala, 2005; and Wilson, Kennedy, Brown, & Draney, 2005.)



[a] Hold volume constant
[b] Hold liquid (water) constant
[c] Hold mass constant
[d] Intuitive Explanations are those that are a student's deeper understanding about WTSF but may not have the appropriate vocabulary.
[e] Problematic Explanations are those where a student uses the correct words (e.g., density), but components of the explanations reflect a more naïve understanding about WTSF.

Figure 2: FAST Buoyancy Progress Variable on Why Things Sink and Float.

(WTSF): Developmental Model Categorizing Students Evolving Conceptions

A second progress variable was constructed to represent the growth in students' skill in reasoning from evidence. It too was based on a five-point developmental trajectory, as shown in Figure 3.

What the Student Can Already Do	What the Student Needs to Improve
Principled Student uses an abstract principle that applies to objects in general.	
Relational Student uses a specific relationship of the form "because X is Y," where the identity of X is made obvious.	To progress to the next level, student needs to use a principle that would apply to objects in general, not just the specific object in their answer.
Unclear Relational Student uses a specific relationship of the form "because X is Y," where the identity of X is not made obvious.	To progress to the next level, student needs to explicitly identify all aspects of the relationship in their explanation.
Experiential Student justifies their answer by appealing to prior experience, having already seen or been told what will happen.	To progress to the next level, student needs to use a relationship to explain their answer, not just evidence to justify it.
Inadequate Explanation Student restates their answer as an explanation, simply asserts that their answer is correct.	To progress to the next level, student needs to understand what evidence is and the relationship between evidence and an explanation.
Off Target Student cannot or does not give an explanation for their answer.	To progress to the next level, student needs to justify their answer in some way.
No Response Student did not attempt to answer.	To progress to the next level, student needs to responds to the question.
Unscorable Student gave a response, but it cannot be interpreted for scoring.	

Figure 3: FAST Reasoning from Evidence Progress Variable

Study Sample

To secure a sample for the study, leaders of professional development networks across the state were contacted to develop a list of potential subjects— middle school teachers who were experienced in the use of hands-on science curriculum, who might be willing to implement the specific 10-12 week FAST unit on which the study was based. Study requirements included participation in a week-long summer institute to

orient the teachers to the curriculum and its associated formative assessments and completion of required teacher and student data over the course of the study. As a result of the recruitment, 13 teachers agreed to participate in the study and attended a five-day summer institute. The summer institute introduced teachers to the FAST unit and its accompanying reflective lessons, including philosophy and underlying learning principles of each; and provided opportunities to for teachers to engage in sample activities and reflective lessons, and to apply the scoring rubrics to student work. Prior to the start and at the end of the institute, teachers also completed pre-assessments of their content knowledge, which mirrored the pre- and post- assessments that their students were to complete.

These 13 teachers represented 13 different middle schools from across the state serving a range of community types, ranging from a private school serving a relatively affluent community, to urban sites serving economically disadvantaged students of color and limited English proficiency, to suburban and rural sites. Of these 13 teachers, eight completed the units and participated in all aspects of data collection; one teacher completed the unit and was observed, but did not submit student data and another teacher completed the unit and submitted student data, but was not observed. Three teachers failed to initiate or complete the units.

Data on the background and experience of the sample who were observed (n=9) suggest they were a highly capable group. Of those who completed the unit, all but one had an undergraduate major in science or science education and the majority had at least a master's degree in science or science education. Years of teaching experience and teaching science ranged from 6 to 30 years, and six of the nine teachers were female. It is of interest that of the teachers who participated in the summer training but did not implement the FAST unit, one had limited formal background in science, and the other had only one year's teaching experience.

Tables 2 and 3 summarize information about teachers' assessment experience and practices, which was collected prior to project training. Teachers were queried about their knowledge of the FAST unit content (mass, volume, density, relative density, etc.); their knowledge and implementation of various kinds of assessments; the frequency of such implementations; and the extent to which they used results to take action at the individual, subgroup, or whole class levels and/or to refine curriculum for subsequent years. Table 2 shows the reliabilities of scales composed from teacher survey responses, which while moderate are reasonable given the small numbers of items that constituted each scale and the small number of respondents.

The data in Table 3 suggest that the sample view themselves as highly knowledgeable about the scientific concepts addressed in the FAST curriculum and as practiced in assessment use. These data provide additional evidence that our sample do not represent typical teachers. Interestingly, the differences between self reported frequencies of assessment implementation and use of results suggest that respondents do not consistently take action based on their assessments.

Table 2.

Summary of the Alpha Coefficients on Teacher Assessment Practice Scales and Subscales

Scale	Alpha	Items (n)
Teacher assessment practice	.73	6*
Assessment implementation:	.69	4*
Frequency of using classroom assessments	.60	7
Use of assessment practices	.86	5
Use of assessment at multiple levels of analysis	.68	4
Knowledge of assessment	.77	12
Knowledge of FAST content	.72	7

* Scale items consist of composite variables created from the subscales.

Table 3.

Summary of the Scores on Teacher Assessment Practice Scales

Scale	Mean	SD	n
Teacher assessment practice	3.21[a]	.35[a]	9[a]
Assessment implementation:	3.02[a]	.37[a]	9[a]
Frequency of using classroom assessments	2.87[a]	.43[a]	9[a]
Use of assessment practices	2.96[a]	.72[a]	9[a]
Use of assessment at multiple levels of analysis	2.69[a]	.53[a]	9[a]
Knowledge of assessment	3.75[b]	.33[b]	9[b]
Knowledge of FAST content	3.95[b]	.46[b]	9[b]
Classroom assessment training in the past three years	3.57[a]	.53[a]	9[a]

Note: All items are on a 5-point scale. For "a" items, 1 = never, 5 = daily.
For "b" items, 1 = not at all knowledgeable, 5 = very knowledgeable.

A summary of the student demographics in our sampled classrooms is shown in Table 4. The data show a wide range of class sizes among observed teachers, ranging from 21 to 40 students. Representation of low SES and minority students, ranging from 0 to 72% and from 5 to 90%, respectively, similarly shows a moderately diverse sample. Few students were classified as limited English proficient or as special needs students.

Table 4.
Class Level Summary of Student Demographics[1]

Background Variable	Mean	SD	Minimum	Maximum
Class Size	30.13	5.67	21	40
Percentage of female students	49.13	15.33	33	76
Percentage of minority students (not White)	36.38	32.07	5	90
Percentage of low SES students[2]	15.13	27.98	0	72
Percentage of students classified as LEP[3]	4.38	6.39	0	15
Percentage of special needs students[4]	3.00	5.66	0	14

[1] The means of the average class composition for each of the 8 sites.
[2] Low SES indicated through participation in the free lunch program.
[3] LEP: Limited English Proficiency classification.
[4] Students classified as having physical and/or mental disabilities.

Measures of Assessment Implementation and Use

Measures of teachers' implementation and use of the reflective lessons were derived from multiple sources. These included classroom observations as the reflective lessons were implemented, pre- and post-observation interviews, and web-based teacher reflection logs.

Observations of reflective lessons. Classroom observations of the reflective lessons involved multiple protocols. During the period of observations, observers took running notes and during periods of whole class discussion, tallied the nature of teacher behavior and each instance of individual student questioning and/or responding.[3] Observers also tallied student engagement during small group activity. After the completion of the observation, observers summarized their observations in a series of rating scales, with justification, that were adapted from Horizon Research, Inc. 2002-2003 Classroom Observation Protocol (Horizon Research, Inc. 2002-2003 Core Evaluation Manual: Classroom Observation Protocol, September, 2002).

Teacher behaviors. General categories were used to classify teacher behaviors: (a) directions; (b) direct instruction/elicit information; (c) elicits students' conceptual understandings; (d) provides feedback; and (e) other non-instructional. Specific

[3] Observations were tape recorded, so that observers could use the tape to complete the teacher tallies. Student engagement tallies were done in real time.

behaviors were tallied within each of these categories. For example, the direction category included giving or clarifying procedural directions, checking for procedural understanding, and reviewing criteria. The second category included instances of direct instruction, elicitation of prior knowledge and/or specific observations, teacher demonstration—none of these activities called on students' deeper understanding. The latter types of activities were reserved for category three, which included asking for ideas and explanations; drawing out alternative conceptions, asking students for synthesis and/or to draw connections; and probing for evidence and/or reasoning. Feedback categories ranged from "right/wrong" to substantive, descriptive feedback and building on students' responses with targeted questioning. Other included housekeeping, disciplinary issues, and other interruptions to instruction.

Student engagement during whole class activities. During whole class activities, student engagement was classified by source of question initiation and response and by the number of different individuals responding. Categories included: teacher initiates, student responds; student initiates, teacher responds; student initiates, student responds; teacher initiates, teacher responds; teacher initiates, teachers repeats questions.

During small group activities, students' engagement in each group was noted at the instance of observation. Categories included talk/procedural; talk/conceptual; manipulation only; read/write; look/listen on task; disengaged; waiting for teacher.

The summary analytic protocols included synthesis ratings of multiple scales and ratings of multiple indicators within each scale. Observers were to use their indicator ratings to inform their synthesis ratings and to provide specific evidence, including examples and/or quotes, to justify each score. The scales included fidelity of FAST implementation, evidence of teacher content knowledge, nature of teachers' formative queries, lesson arrangement, student engagement and questioning strategies, and overall synthesis of the quality of assessment practice observed. Most of the scales used a 1-5 rating continuum, with 1 indicating not at all in evidence and 5 indicating consistently in evidence.

Observers were trained to use the protocols in a series of meetings. Category and coding definitions were operationalized and applied to video clips. Repeated cycles of practice, discussion, and coding refinement occurred until consensus was reached. Reliability checks were conducted by pairing raters in a sample of classrooms.

Each teacher was observed at least once, during Reflective Lesson 7, a mid-point in the FAST curriculum. Most teachers were observed on multiple days and for multiple reflective lessons. Teachers were also were interviewed informally before and after each observation to ascertain how the units and their assessments were going, what challenges had arisen, special plans for the period of observation, and sense-making and reflections on the what had been observed.

Teacher logs. After each investigation, teachers were asked to complete a web-based survey form in which they indicate the proportion of their students who fell into each progress level on the content and reasoning progress guides/rubrics and indicated the sources of evidence they had used in assigning these ratings. For the Reflective Lessons, teachers also were asked to indicate what they thought their students had learned from engaging with the Reflective Lessons and how, if at all, their understandings of the curriculum and student understandings had been influenced by the Reflective Lessons.

Final teacher interview. The final interview asked teachers about unit implementation in terms of the amount of time the unit took to complete, how it worked with students, and changes in the curriculum and assessments that may have been implemented. The interview questions particularly probed teachers' use of the Reflective Lessons, the strategies and understandings which may have emerged from their use of the progress guides, and the effects of the study on their content knowledge, teaching practices, and use of assessment.

Student Performance and Demographic Data

The study included multiple sources of student data. These included specially designed pre- and post-assessments, student performance on the reflective lessons themselves, and available archival data.

Pre and Post Assessments. See Wilson, Kennedy, Brown, & Draney (2005) in same session.

Student performance on reflective lessons. Teachers submitted their students' responses to the reflective lessons. RL4, RL7, and RL10 were centrally scored by CAESL researchers. Each was double scored based on the project progress guides, with each assessment receiving a score on WTSF (content) and a score for reasoning. (See Wilson, Kennedy, Brown, & Draney [2005]). Analyses of these data, and particularly the relationship between teachers' scores and central research scores, are not yet completed.

Archival data. Teachers submitted individual data on their students' performance on California's accountability testing. These data are not yet complete. Moreover, teachers provided demographic information for each student, including gender, ethnicity, language status, special education status, and free lunch status.

Analysis Plan

Both quantitative and qualitative methods were used to analyze available data. Our approaches are described in the sections that follow/

Case summaries of teacher assessment implementation and use. Observers synthesized all available data in case summaries for each teacher. A common format was used in the development of each case, and observers agreed on the data sources that were to inform each section, including particular elements from the field notes, observation protocols, after observation synthesis ratings, interviews, teacher logs, and available context data. Each case addressed: Contextual information including nature of school context, student demographics and teacher background; descriptive chronology of observed classes; ratings and evidence-based perspectives on: quality of FAST implementation; extent of teacher content knowledge, nature of queries and question types; extent and nature of feedback; quality of student engagements; teachers' analysis and use of data; and overall quality of teachers' assessment practices. The CAESL researchers came together repeatedly to discuss and compare the nature of their teachers' practices in each area and to discuss different elements of quality within each.

For the purpose of the initial quantitative portion of the study, the number of teachers in our sample pushed us to reduce our qualitative findings to a single indicator of the quality of assessment practices. Though discussion, comparison and debate, the researchers agreed on the elements of their overall ratings of assessment practices and the characteristics of each teacher that placed him/her in a particular overall rating category, the possible values of which included:

Level 1: Ineffective Formative Practice
Level 2: Elements of Effective Formative Practice
Level 3: Beginning Stages of Formative Practice – Low (2.7)
Level 3: Beginning Stages of Formative Practice – Solid (3.0)
Level 3: Beginning Stages of Formative Practice – High (3.3)
Level 4: Accomplished, Effective Formative Practice
Level 5: Exemplary Formative Practice

As we discuss later, one challenge the researchers faced was to clearly differentiate formative assessment quality from other aspects of effect pedagogy.

Hierarchical linear modeling (HLM) analyses of the relationship between formative practice and performance. Hierarchical linear modeling (HLM) was employed to examine the influence of both student and teacher level variables on student performance. While HLM provides predictive values (coefficients) similar to a regression analysis, a key difference is that it allows the study of multiple levels of analyses, estimating between-group differences (Raudenbush & Bryk, 2002; Seltzer, 1994). Another benefit of HLM is that it is able to reliably perform analyses with unbalanced data as well as small sample sizes (Raudenbush & Bryk, 2002; Seltzer, 1994). This study examine the relationship among and between student (Level 1) and classroom (Level 2) level variables and student post-test performance on the Buoyancy and Reasoning progress variables.

Level 1, student variables, included student demographics (ethnicity, free lunch status, language status, disability status) and entering capacity, scores from the pre-tests. Level 2 variables included teacher and assessment practice indices. The current analyses only included data from the seven teachers for whom we had complete data at the time of the analysis. (Future analyses may add additional teachers). In addition to looking at the relationship between measured practices and student learning, the HLM analyses were also used to identify teachers whose students showed substantially more or less learning relative to the total sample.

Results

We provide below a synthesis of our descriptive results, followed by the results of the HLM analyses. Despite the quality of the FAST curriculum and embedded assessments, the detailed implementation guidance, and the strong content and educational experience of participating teachers, results of the study show considerable variability and leave a number of questions unanswered about the nature of teachers'

use of quality of assessments tools and their related impact on student progress and achievement.

Descriptive Results

Content background. Consistent with the self-reports from the teacher surveys, observations suggested that all teachers displayed an understanding of the core concepts underlying the FAST unit, that is, of mass, volume, density, and relative density and their relationship to buoyancy. Teachers displayed such understanding through dialogue with students during presentations and class discussions; they appeared confident with the content and were able to make appropriate connections to ideas in prior lessons and in real-world contexts; and generally they were able to engage students intellectually with the relevant ideas grounding specific lessons.

In one FAST classroom, in response to a student question, a teacher generated an "on-the-fly" demonstration to explain the relationship between mass and volume. In another classroom, a teacher opened a lesson with a clarification of the definition of buoyant forces. Both instances revealed the depth of teacher content knowledge, a critical aspect of effective instructional and assessment strategies.

Quality of FAST implementation. In noting the extent to which the implementation of the formative assessments appeared consistent with the FAST philosophy and its inquiry-oriented and student-centered instructional strategies, observers found that all teachers at least attempted to implement the curriculum and assessment activities as specified in guidance documents, using student explorations, small group interactions, and whole class discussions as recommended. Teachers showed respect for students' ideas and questions, students seemed to respect each others' contributions, and most students appeared at least somewhat engaged with the lesson content.

Implementation, however, varied considerably with regard to lesson pacing and the extent to which implementation mirrored the spirit—and not solely the superficial surface features—of FAST. For some teachers and classrooms, the pace was appropriate to students' needs and the intent of the curriculum, and most all students appeared engaged in the lessons throughout. Multiple responses were elicited to the same question querying conceptual understanding to reveal multiple student conceptions, students had the opportunity to listen to various arguments and explanations of specific phenomena, and the teacher displayed confidence in his/her management of the class and implementation of the lessons and assessment. In these classrooms, pacing of the

lessons and assessment strategies were varied—moving among and between whole class, individual, small group, and pair share activities—and appropriate to the level of class and the kinds of ideas and understandings expressed by students. Students were provided with multiple opportunities to ask questions, share information and participate in a variety of learning situations. The teacher encouraged students to think more deeply about their responses, to use other students as resources for additional information, and to listen carefully to predictions and ideas that were shared.

In classrooms where pacing and engagement were high, teachers seemed to have special strategies and well understood routines for maintaining engagement and assuring accountability. For example, in one classroom the teacher assured that every student committed to a prediction during group discussions by having students stand to show their position; the teacher then could follow up easily with students with opposing positions. Similarly, in another classroom, the teacher had students physically move from one side of the room or the other to show their position. In yet another classroom, students were randomly selected from a "kitty" of names to respond to class discussion questions, assuring student accountability and reducing the opportunity for students to "slip under the radar "of responsibility for listening and understanding during whole class discussions. Several of the FAST teachers had well-established routines for individual and small group participation that were closely linked to pacing of the lessons.

Teachers employed a variety of techniques to assure student engagement during small group work, including pair sharing, fearsome foursome, and other small group interactions. Some FAST teachers alternated pairing strategies, at times requesting that high-performing students share ideas with other like minded colleagues, and alternating these groupings with mixed performance levels. One FAST teacher alternated table groupings on a monthly basis, after a careful review of student performance on both written and whole class discussions. In other cases, however, pacing and student engagement were less successful. While a teacher might pose questions to elicit students' multiple conceptions and indeed successfully elicit alternative explanations, probing was ineffectual or absent in revealing underlying rationales or in helping students to confront their misconceptions. Whole class discussions ended when the lesson period ran out, ending with no resolution of students' conflicting ideas and no attempt at synthesis. While leaving students hanging could be conceived as a reasonable motivation or attention strategy for subsequent

follow up, the latter did not occur. Students who completed their individual work early were left to dawdle until all children were ready to move onto the next activity.

The quality and level of student engagement during small group activity was likewise variable. In the absence of clear norms for the level and quality of interactions and discussion and established routines and expectations, students appeared to find it difficult to engage in sustained activity or in substantive discussion with their peers, unaided by the teacher, and their attention wandered. Even when small groups were engaged, there was a tendency to rely on one or two students within the group to carry the workload.

Observations suggest that teachers with clearly established routines for participation in group work and accountability for thinking and articulating group findings were more successful involving all students in the class discussions, and ensuring that all students had multiple opportunities from which to clarify and deepen their understandings of concepts. As noted previously, various grouping structures, random selection of students to whom questions were addressed, and a host of other classroom management techniques and instructional strategies helped support quality classroom interactions.

Teachers' implementation strategies also varied widely in the degree of structure and scaffolding they employed. Even as they pursued FAST's intent to elicit students' understandings, some teachers used strategies such as sentence starters to prompt student responses. For example, in conducting a discussion of the relationship between mass, volume, and sinking and/or floating, one teacher guided the discussion by progressing through a series of sentences, eliciting reasoning along the way:

Examplar 1: FAST Teacher Scaffolding

1. "As the _____ (increases or decreases), then the _____(increases or decreases)."
2. "The ____ equals _____."
3. If the _____ is (greater or less), the _____ will _____(float, sink, subsurface float."

Another teacher chose to supplement FAST small group discussion questions with more detailed questions to guide students' thinking, as well as to provide smaller chunks for prompting small group discussion, a strategy she felt would help to optimize student engagement.

Distribution and nature and teacher questioning. In recording the teacher behavior during whole class time, the observers focused on the following general categories:

1. Giving directions
2. Providing direct instruction, eliciting factual information
3. Engaging students in questions involving conceptual understanding
4. Providing feedback
5. Non-instructional business

Categories three and four present the heart of FAST and the core goals of the reflective lessons that were observed. Yet teachers varied widely in their relative attention to these various activity types. The contrast between the most and least interactive teachers was striking. For two of the observed teachers, more than half of their observed interactions reflected category three—using questions to elicit and probe their students' ideas; asking for students to generate explanations, hypotheses and predictions; and probing for meaning and evidence.

For example, for one teacher this meant eliciting multiple responses to the same question, asking students for evidence to support their predictions and explanations, requesting that they compare their ideas and predictions to other students' ideas, and to provide evidence of a principle or concept previously discussed or presented. In this classroom, students were regularly asked to make connections to other ideas and concepts from prior investigations. At the other extreme of the questioning continuum, virtually none of a teacher's whole class interactions focused on students' conceptual understanding. Instead, nearly all interactions involved categories one and two, providing directions and direct instruction/eliciting factual information.

Table 5.

Relative Frequency of Categories of Teacher Behavior

Type of Interaction*	**All FAST Ts**	**Range**
Giving directions	27%	7% - 57%
Providing instruction, eliciting information	22%	3% - 36%
Engaging student in questions regarding conceptual understandings	37%	0% - 58%
Providing feedback	9%	0% - 26%
Non-instructional business	5%	2% - 28%

*Note: percentage of interactions allocated to each category during RL7.

Interestingly, Table 5 shows that FAST teachers spent the majority of their time engaged in the types of interactions most critical to the development of student learning and ideas, but relatively little time on providing feedback to students (as will be discussed in subsequent paragraphs).

Regardless of question type, by far the predominant style of questioning during whole class interaction for all CAESL/FAST teachers was "teacher questions-student responds," with limited instances of students raising questions of the teacher or of peers or students responding to student-raised questions. Most all teachers used wait time to encourage student responses and managed to distribute questions over a number of students.

Use of feedback**.** Use of feedback also varied considerably, but surprisingly in most cases was relatively rare. At one end of the feedback continuum, a teacher who posed no questions involving conceptual understanding provided no feedback at all to students during whole class interactions. In contrast, about a quarter of another teacher's interactions involved feedback, but nearly half of these simply noted whether students were "right" or "wrong" and failed to provide the descriptive feedback or substantive follow-up that has been associated with learning increases (Kluger & DeNisi, 1996; Serna et al, 1992).

Moreover, even as feedback was relatively rare, rarer still was it for teachers to go beyond traditional types of feedback—such as "good", "ok", "what do you mean?"—to the kind of feedback that builds on students' responses by targeting questioning to confront misconceptions or to draw students into a sophisticated, conceptual level of discussion and hopefully, understanding. In the best examples of effective feedback, teachers made clear their expectations for quality performance, listened carefully to students' ideas and responded with targeted probing and prompting during whole class or small group discussion. Informal strategies for effective, quality feedback including roaming the classroom, systematically gathering information by reading student work, listening to their interactions, and at times answering individual questions—and then providing specific, targeted information to aid student performance.

Overall Judgments of Assessment Quality

The overall scoring of FAST assessment quality represents an integration of the qualitative aspects of the observations, ratings from the various scales, and pre/post teacher interviews. As described previously in this paper, most FAST teachers implemented the curriculum in ways consistent with the guidelines and training, possessed reasonable knowledge, provided multiple opportunities in which students could share ideas and information, and in general, administered and used data generated by the Reflective Lessons, in mostly informal ways.

In terms of providing students with direct feedback on the written aspects of their performance on the reflective lessons, targeted, specific responses to students were limited. Most teachers claimed to at least glance through all the student work to see how students were doing on the two learning trajectories— WTSF and Reasoning—and to look for patterns. However, RL responses were infrequently if ever returned to students in a timely basis or with written feedback. An exception was a teacher who had an aide to help with record keeping and considered the Reflective Lessons as a grading opportunity and thus assigned grades to each piece. For the majority of teachers, it appeared that they did not evaluate individual students' Reflective Lessons on a timely basis. For many, interpretation or "scoring" of student work did not occur until after the unit was completed, and analysis of patterns and implications was done informally, at best. Such inattention to timely feedback to students may have been a function of the study research procedures, in that responses to Reflective Lessons were to be copied and returned to the researchers, and/or the fact that many teachers

reported being careful not to "contaminate" the data they were submitting for the study.

Formative use of assessment results. It is difficult to use assessment results to modify current instruction in the absence of the existence of such results, and indeed the latter was generally the case for teachers in this study. However, as noted above, a number of teachers commented that informally, they noticed patterns of students' (mis)understandings during whole class discussions and through observations of students' interactions and work during class. Similarly, they also mentioned quickly "thumbing through" student responses to find patterns in what students understood about sinking and floating relative to the learning trajectories and to discover what gaps and problems existed in student understanding. One teacher quickly sorted students' responses into piles reflecting different learning issues and then planned subsequent instruction and grouping around those results. For the most part, however, while teachers claimed to provide group feedback on performance and to deal with misunderstandings in the course of subsequent instruction, they were somewhat reluctant to go back and reteach or involve students in additional activities to directly address their misunderstandings, even though the FAST reflective lesson materials included specific activity suggestions for remedying particular misconceptions or gaps in understanding. Three teachers were exceptions. They did, after reviewing student work, use the results to reteach or review particular concepts or ideas. But reteaching or formal reviewing of content was relatively rare, perhaps because of a perceived (and partially true time constraint), teachers felt pressure to move ahead in the unit.

While comprehensive, systematic and targeted formative use of assessment results was relatively infrequent, teachers were very positive and enthusiastic about the value of the progress guides and the developmental continuum of understanding represented by the progress guides. Teachers appreciated the clarity of expected goals and used the developmental trajectories to focus their instruction, their thinking about student progress, and their informal responses to it. The formal use of formal assessment results may have been limited, but a number of teachers mentioned internalizing the progress trajectories, and using them informally to integrate and gauge students' learning and to pose questions and next steps—even as those next steps might have been relatively minor alterations to what was originally planned.

Overall judgments of quality of formative practice. In terms of overall quality of formative assessment practice, the majority of teachers who were observed (n=8) were judged to be in the beginning stages of effective formative practice (a "3" on the five-

point scale described above), with three of the five so judged characterized at an advanced level of beginning practice. Two of the observed teachers were judged to have achieved accomplished formative practice (a "4" on the five-point scale), and only one teacher was judged a "2," the teacher who generally followed the activity sequence of FAST and the reflective lesson, but failed to engage students on substantive dialogue about their conceptions of buoyancy and why things sink and float.

Teachers who were judged as "4" ostensibly were able to orchestrate all aspects for the formative assessment. They not only posed the right questions, but they had effective routines in place to help assure that most all students were asked to and were engaged in responding, and they targeted questions to elicit and encourage students to confront differing conceptions and facilitated closure in whole class discussion. They also had strong routines and varied activities to help assure that students were engaged in substantive dialogue during small group or pair work. These teachers also were more proactive than the others in soliciting feedback from classroom interactions, observations, and student work and reflecting on how instruction might be modified to better support student learning.

Teachers who were judged a "3" were a mixed set. A common issue was superficial implementation—teachers asked the "right" questions to solicit student conceptions, and may have even probed for explanations and evidence. Even so, they were not able to involve most of students in thinking deeply about their understandings or in comparing and contrasting different conceptions and tended to short change such discussions without bringing them to much closure. Routines for assuring student engagement during small group work were a particular weakness, as has been noted previously. In these classrooms, students were reluctant to discuss with one another, without direct prompting or attention of the teacher. If teachers struggled in probing students' responses, encouraging substantive interactions, and uncovering the reasons underlying various conceptions, it should come as no surprise, that students too would have difficulty. In many cases, however, students instead simply opted out, by not participating or attending.

Results from HLM Analyses

As noted above, hierarchical linear modeling was performed to examine the influence of both student and classroom level variables on student post-test scores. The small sample size, however, severely limits these analyses. Student level variables included indicators for ethnicity, SES, language status, special needs status, and pre-test

performance. Test scores are expressed as probability ratios, with higher values indicating a greater chance that a student possesses a high level of understanding relative to the progress variables representing the major content and inquiry process progress variables in the study. If the estimate is high for the Reasoning variable, then we would say that the student uses more sophisticated reasoning in their justifications.

Table 6 summarizes data on students' pre- and post-test performance for the seven teachers for whom we had complete data. The data in Table 6 suggest that students developed better understandings of buoyancy and improved their skill in reasoning from evidence as a function of the FAST unit, in that post-test scores are significantly higher than pre-test scores in both areas, the WTSF gains are 1.12 and those for reasoning are .93. Teaching and learning in the FAST classroom, supported by a quality curriculum and assisted with access and use of quality assessments to carefully monitor student progress resulted in students learning and developing their understandings of mass, volume, and relative density.

Table 6
Summary of Student Progress Variable Scores

Reflective lesson	Mean	SD	N
Why things sink and Float (pre-test)	-.35	.66	207
Why things sink and Float (post-test)	.78	.66	201
Reasoning (pre-test)	-.41	.87	207
Reasoning (post-test)	.52	.61	201

HLM analyses first looks at the variation explained by the first level variables—in this case individual student variation in demographics and pre-test performance. Then, assuming there is remaining variance to explain, the analyses examine second level variables—in this case teacher variables. The first level modeling explained only .3 of the variance in post test performance, leaving the majority of variance unexplained, even with the inclusion of pre-test performance, and suggesting a poor model fit. Results from adding the teacher level variables were even more paltry, accounting for only 6% of the variance in student scores on WTSK and 3% on the reasoning variable. The disappointing results in Table 7 are a function of a number of limitations of this

study: the small number of teachers; because of missing data the relatively small numbers of special needs students (n=6), low SES (n=35), minority (African American or Latino students, n=35); and the relatively small variance in our assessments of the quality of teachers' practices—all teachers for whom we had student performance data were rated overall as "3" or "4". The data in Table 7 do show, however, that some of the pre-existing teacher background variables did pop out as marginally significant, but given the aberrations of the data, these coefficients warrant little interpretation.

Table 7
Summary of Coefficients for the Two-Level Hierarchical Modeling of the FAST Post-tests on "Why things sink and float" and "Reasoning"

WTSF	Level 1 intercept	Special needs	SES	Minority	Pre-test	Teacher Between-Group Variance
Intercept	.75***	-.80*	-.14	-.32*	.42***	.06***
Frequency using assessments Slope	-.40	1.82*	.88+	.50	-.27	(.30)
Intercept	.75***	-.04	-.08	-.24+	.38***	.09***
FAST Knowledge Slope	-.16	1.32+	2.84	.25	.01	(.31)
Intercept	.75	.12	-.07	-.24+	.39***	.06***
Multiple levels of analysis Slope	-.29	-3.77+	1.56	.27	-.17	(.31)
Reasoning						
Intercept	.54***	-1.06**	.01	-.12	.23***	.03**
Frequency using assessments Slope	-.14	1.86*	.14	-.09	.05	(.33)
Intercept	.54***	-.21	-.02	-.13	.25***	.03**
FAST Knowledge Slope	-.13	1.43+	2.60	-.02	.10	(.33)
Intercept	.54***	-.04	-.01	-.13	.22***	.03**
Multiple levels of analysis Slope	.07	-4.11+	1.50	.01	.07	(.33)

+ $p < .10$
* $p < .05$
** $p < .01$
*** $p < .001$

The simple results by the teacher in Table 8, controlling for student demographic characteristics and pre-test performance are telling. Consistent with the HLM findings, there is relatively little adjustment in scores based on background and pre-test

performance. In both cases, one teacher jumps out as showing relative higher post-test performance on WTSF and another shows relatively low performance. Most are bunched up in the middle range. It is of interest that Teacher #3 was rated as a solid 3 on overall quality of formative assessment. While she "listened" and observed what students were learning, because of large class size, she provided few opportunities for classroom discussion and little probing of students' understanding. She was amongst the most highly structured of the observed teachers and treated the reflective lessons as a grading opportunity. Her class was located in a private school serving a middle class population. Teacher 7, who jumps out as low on both WTSF and Reasoning, was rated between a high 3 and low 4 in overall quality of formative assessment. She was one of the few teachers who engaged in re-teaching and she taught a relatively privileged group of students.

Table 8:
Comparison of the Mean Post Test Scores by Teacher for Adjusted Scores Controlling for Background Variables [1]

Reflective lesson	Adjusted Mean[2]	Unadjusted Mean[3]
Why Things Sink and Float		
Teacher 1	.84	.73
Teacher 2	.53	.58
Teacher 3	1.31	1.33
Teacher 4	.69	.81
Teacher 7	.38	.46
Teacher 8	.68	.52
Teacher 10	.78	.75
Post-test score gap between teachers	.93	.87
Reasoning		
Teacher 1	.45	.42
Teacher 2	.72	.73
Teacher 3	.47	.47
Teacher 4	.52	.58
Teacher 7	.28	.33
Teacher 8	.48	.40
Teacher 10	.87	.85
Post-test score gap between teachers	.59	.52

[1] Controlling for pre-tests, special needs, SES, and minority status.
[2] Class mean post-test score.
[3] Class mean post-test score adjusted for background variables.

We are continuing our analyses to detect qualities of practice that make a difference. However, current findings suggest the difficulty of detecting effects of formative assessment in complex environments with small *n*s and are a good reminder of the number of variables and circumstances that make a difference in student learning.

Conclusions

One of the goals of this study was to better understand the ways in which teachers use quality tools to support and promote student learning and achievement. To do so, teachers must collect and use assessment data consistently and systematically to inform both the nature and types of feedback they provide to students, and to make decisions about how and in what ways to guide instruction. The current study, while

disappointing in many respects—after all, shouldn't strong instructional leaders, using good curriculum and assessment tools, provide a reasonable context in which to determine the way/s in which quality assessment practices are employed, and provide reasonable ways in which to differentiate more effective assessment practice from less successful practices?—raises important questions.

So where do we look to explain differences in performance, differences in what students learned about "why things sink and float" from the FAST curriculum? It is clear that effective formative assessment practice is not easy to encapsulate. There is more to successful teaching and learning than simply administering assessments, scoring the assessments and sending data to researchers. Asking the right questions, probing for explanations and evidence is insufficient. There is no simple cookbook to achieve effective practice. Learning must be orchestrated in complex ways that bring together a variety of teacher expertise—strong content knowledge, sophisticated pedagogical knowledge and strategies, effective assessment, and strong routines and norms for student engagement. Researchers, who are prone to show disdain for teachers' concerns for discipline and classroom management, too often ignore the latter. Yet without student engagement—time on task as we used to call it—the best "methods" cannot yield promised dividends in students' learning.

The challenge of effective assessment for learning is daunting indeed. Even in the hands of highly qualified, well-trained, sophisticated teachers, with a well-structured curriculum, quality assessment tools must be used in quality ways to make a difference. Absent from the present study were many elements of what the literature currently describes as "learning communities": In this study, teachers essentially worked in isolation, without the benefit of collaboration or the opportunity to look at student work in the context of the lessons. Absent too, was the freedom (real or perceived) for teachers to alter the curriculum or reflective lessons to better or more appropriately fit student needs. Finally, other CAESL assessment projects are finding the importance and need for teachers to better understand how and in what ways to analyze student performance in a more systematic manner based on performance of specific subgroups, and then devise or employ specific instructional strategies to address alternative conceptions.

What else did we learn from this study?

1. To state the obvious teaching a new curriculum, and using new assessments requires time, energy and a great deal of trial and error, even for

accomplished, experienced, talented and knowledgeable teachers. Many FAST teachers planned to teach the unit again to a different class, and to use the knowledge and experience they gained from their pilot experience

2. The CAESL tetrahedron is a potentially useful tool for understanding classroom assessment practices, but much work remains to better understand the specific dimensions of what constitutes quality tools and quality use.
3. The process of looking at student work in a timely manner, scoring the work and providing timely, meaningful feedback to students is complicated and potentially requires more specific, cyclical, scaffolded learning experiences for teachers. Scoring assessments in a facilitated training session is the beginning of the process, but repeated opportunities to engage in the kinds of conversations and conceptual thinking necessary to understand student work is needed.

Finally, the study provides food for thought about the research methods needed to study teachers' assessment practices and the complexity of assessing their effects on student learning. On the one hand, our study suggests that effective formative assessment is a highly interactive endeavor, involving the orchestration of multiple dimensions of practice, and demands sophisticated qualitative methods for study. On the other, detecting and understanding learning effects in small samples, even with the availability of comparison groups, poses difficulties to say the least. Our search continues.

REFERENCES

Ayala, C. (2005). *On the development of CAESL/FAST Reflective Lessons.* Paper presented at the annual meeting of the American Educational Research Association. April 2005, Montreal Canada

Bell, B., & Cowie, B. (2001). *Formative assessment and science education.* Dordrecht: Kluwer Academic.

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

Black. P., & Wiliam, D. (1999). *Beyond the black box. Assessment for learning: Beyond the black box.* Cambridge, UK: University of Cambridge, School of Education. [Assessment for Learning: Beyond the Black Box is available to download as a PDF file.]

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1*(2), 1-12. (ERIC Document Reproduction Service No. ED 053419)

Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24,* 355-392.

Gitomer, D. H., & Duschl, R. (1995). Moving toward a portfolio culture in science education. In S. M. Glynn & R. Duit (Eds.), *Learning science in the schools: Research reforming practice* (pp. 299-325). Mahwah, NJ: Lawrence Erlbaum Associates.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18,* 519-521.

Heritage, M., & Yeagley, R. (2005). Data use and school improvement: Challenges and prospects. *Yearbook of the National Society for the Study of Education, 104,* 320-339.

Herman, J., & Gribbons, B. (2001). *Lessons learned in using data to support school inquiry and continuous improvement: Final report to the Stuart Foundation.* Los Angeles: University of California, Center for the Study of Evaluation.

Horizon Research, Inc. (2002). *2002-2003 Core evaluation manual: Classroom Observation Protocol, September, 2002.* Chapel Hill, NC: Author.

Killion, J. (1999). *What works in the middle: Results-based staff development.* Ann Arbor, MI: National Staff Development Council. Retrieved 6 November 2007 from http://www.nsdc.org/midbook/

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

National Research Council. (2001a). *Classroom assessment and the National Science Education Standards.* Committee on Classroom Assessment and the National Science Education Standards. J. M. Atkin, P. Black, & J. Coffey (Eds.). Center for Education, Division of Behavior and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2001b). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. In J. Pellegrino, N. Chudowsky, & R. Glaser (Eds.). Board on Testing and Assessment, Center for Education, Division of Behavior and Social Sciences and Education. Washington, DC: National Academy Press.

Pauls, J., Young, B., & Lapitkova, V. (1999). Laboratory for learning. *The Science Teacher, 66*(1), 27-29.

Plake, B. S., & Impara, J. C. (1997). Teacher assessment literacy: What do teachers know about assessment? In G. Phye (Ed.), *Handbook of classroom assessment* (pp. 53-68). San Diego, CA: Academic Press.

Pottenger, F., & Young, D. (1992). *The local environment: Fast 1 Foundational Approaches in Science Teaching.* Honolulu: University of Hawaii at Manoa, Curriculum Research and Development Group.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Rogg, S., & Kahle, J. B. (1997). *Middle level standards-based inventory.* Oxford, OH: Miami University.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 1–25.

Seltzer, M. H. (1994). Studying variation in program success: A multilevel modeling approach. *Evaluation Review, 18,* 342–361.

Serna, L., Schumaker, J., & Sheldon, J. (1992). A comparison of the effects of feedback procedures on college student performance on written essay papers. *Behavior Modification, 16*(1), 64-81.

Shavelson, R., Stanford Educational Assessment Laboratory (SEAL), and Curriculum Research & Development Group (CRDG). (2005). *Embedding assessments in the FAST curriculum: The romance between curriculum and assessment. Final report.* Stanford, CA: Stanford University, Stanford Educational Assessment Laboratory.

Shepard, L. A. (2001). The role of classroom assessment in teaching and learning. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 1066-1101). Washington, DC: American Educational Research Association.

Skinner, B. F. (1953). *Science and human behavior.* New York: MacMillan.

Skinner, B. F. (1960). Teaching machines. In A. A. Lumsdaine & R. Glaser (Eds.), *Teaching machines and programmed learning* (pp. 137-158). Washington, DC: National Education Association.

Stiggens, R. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83,* 758-765

Tamir, P., & Yamamoto, K. (1977). The effect of the junior high 'FAST' program on student achievement and preferences in high school biology. *Studies in Educational Evaluation, 3*(1). 7-17.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction.* Chicago: University of Chicago Press.

U.S. Department of Education, Office of Educational Research and Improvement, System of Expert Panels, Mathematics and Science Educaiton. (2001). *2001 Exemplary and Promising Science Programs—Foundational Approaches in Science Teaching (FAST)*. Washington, DC: Author. Retrieved 6 November 2007 from http://www.ed.gov/offices/OERI/ORAD/KAD/expert_panel/fast.html

Wilson, M., Kennedy, C., Brown, N., & Draney, K. (2005, April). *Using progress variables and embedded assessment to improve teaching and learning.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*, 181-208.

Young, M. J. (1993). *Quantitative measures for the assessment of declarative knowledge structure characteristics.* Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh, PA.