CRESST REPORT 740

Joan L. Herman
Kilchan Choi

FORMATIVE ASSESSMENT AND THE IMPROVEMENT OF MIDDLE SCHOOL SCIENCE LEARNING: THE ROLE OF TEACHER ACCURACY

AUGUST, 2008

National Center for Research on Evaluation, Standards, and Student Testing

Graduate School of Education & Information Studies
UCLA | University of California, Los Angeles

Formative Assessment and the Improvement of Middle School Science Learning: The Role of Teacher Accuracy

CRESST Report 740

Joan L. Herman and Kilchan Choi
CRESST/University of California, Los Angeles

August 2008

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
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported in part under grant ESI-0119790 from the National Science Foundation to WestEd for the Center for the Assessment and Evaluation of Student Learning (CAESL).

The findings and opinions expressed in this report do not necessarily reflect the positions or policies of the funding agency, National Science Foundation.

FORMATIVE ASSESSMENT AND THE IMPROVEMENT OF MIDDLE SCHOOL SCIENCE LEARNING: THE ROLE OF TEACHER ACCURACY[1]

Joan L. Herman & Kilchan Choi
CRESST/University of California, Los Angeles

Abstract

This article articulates a framework for examining the quality of formative assessment practice and provides empirical evidence in support of one of its components. Based on a study of middle school science, the study examines the accuracy of teachers' judgments of students' understanding and the relationship of such accuracy to middle school students' learning. Analyses within and between teachers show a consistent, positive relationship between teacher accuracy and student learning. Study results lend support for the power of assessment in improving student learning and also suggest some potential challenges in assuring quality formative assessment practice.

Introduction

Assessment in the service of learning is a precept with longstanding appeal. The idea traces its roots to the beginning of the field of educational measurement itself (see, for example, Thorndike, 1918), is visible in the growth of criterion-referenced curriculum, instruction and assessment (Tyler, 1948; Glaser, 1963), is evident in mastery learning (Bloom, 1968), assessment driven instruction (Popham, 1987), and suffuses recent state and federal policy and its continuing attention to standards-based testing and accountability (Resnick & Resnick, 1992; Perie, Marion & Gong, 2007). Today, burgeoning interest in formative assessment reflects the view that the most powerful use of assessment occurs hand-in-hand with classroom teaching and learning (Black and William, 1998; Bloom, 1968; Black & Wiliam, 2004; Sadler; 1989; Shavelson, 2006; Shepard 2005).

Yet, while the potential of formative assessment is rich, particularly given the strong effect sizes found shown by Black and Wiliam's meta-analysis (1998) and the special impact it showed for low ability students, the current reality of teachers' assessment practices is less so. Available teaching materials lack the types of systematic and sensitive assessments that teachers and students need to spark and make visible students' thinking or to provide

[1] We also are grateful to our CAESL colleagues whose participation in the underlying study made possible these data: Steve Schneider, Rich J. Shavelson, Mark Wilson, Kathleen Kennedy, Mike Timms, Ellen Osmundson. Thanks to Sam Nagashima for his contributions to the study analyses and to Ellen Osmundson for her helpful review and feedback.

feedback and guide subsequent action. Moreover, in spite of federally mandated demands on teachers and schools to regularly and consistently assess student progress, educators have limited background and capacity to develop and use assessment in general (Heritage & Yeagley, 2005; Herman & Gribbons, 2001; Plake & Impara, 1997; Shepard, 2001; Stiggins, 2005), and the nature and quality of teachers' formative assessment practices specifically has been little studied. Until recently, the scant existing data has demonstrated the uneven quality of teachers' classroom assessments (McMorris & Boothroyd, 1993), even as they have shown that teachers report relying on their own tests and assessments more than external tests for classroom decision-making (Dorr-Bremme & Herman, 1986; Hamilton & Stecher, 2006).

What will it take to move from current reality to quality in formative assessment practice? Available theory and literature seems to pursue two dominant themes, one focusing on the nature of the assessment tools needed to support formative practice and the other emphasizing the process of assessment use. For example, "Knowing What Students Know" (KWSK; National Research Council [NRC], 2001), synthesizing decades of research in cognition, measurement, and psychometrics, clearly articulated the critical role of assessment in advancing learning, the primacy of teachers' classroom assessments in such advancement, and the need to create a new generation of learning-based assessments. Although KWSK conceptualized a powerful model for creating such assessments, it was silent on issues of whether and how these assessments can be used by teachers to improve their students' learning. At the other end of the continuum, "Inside the Black Box" (Black & Wiliam, 1998) and Bell and Cowie's case studies of formative assessment (2001), deal with teachers' use of assessment to promote learning, but largely neglect issues of the quality of the assessments and data used.

Our research attempts to bring both of these perspectives together in a general model defining critical elements in quality formative assessment practice (see also DiRanna et al., 2008; Herman & Heritage, 2007; Herman, Osmundson, Ayala, Schneider, & Timms, 2005). The model specifies quality in goals, assessment tools, interpretation and use of results as key components of assessment that benefits learning. Below, we first lay out the general model and then use data from a small study of curriculum-embedded assessment in middle school science to examine how one of the model's elements, teachers' interpretation of student assessments, may influence student learning.

The data set used here is part of a larger study conducted by the Center for the Assessment and Evaluation of Student Learning (CAESL), funded by the National Science Foundation. The current study builds on an earlier one examining whether and how teachers' use of assessment results is related learning (see Herman et al., 2005). That study found that

teachers made limited use of assessment in explicit planning for subsequent instruction or to provide direct or descriptive feedback to students. Analyses revealed no apparent relationship between the quality of assessment use and student learning, although admittedly the data set was very small and exhibited limited variation in the quality of teachers' assessment use.

Model Components in Formative Assessment

As noted above, and summarized in Figure 1, our perspective combines attention to the quality of assessment (the instruments or tools used) with that to the quality and process of assessment interpretation and use. Figure 1 essentially embeds Knowing What Students Know (KWSK) "triangle," to define assessment quality (NRC, 2001). Advocating the primacy of assessment to benefit student learning, KWSK observed that "Every assessment….. rests on three pillars: a model of how students represent knowledge and develop competence in a subject matter domain; tasks or situations that allow one to observe students' performance; and an interpretation method for drawing inferences from the performance evidence thus obtained" (p. 2).

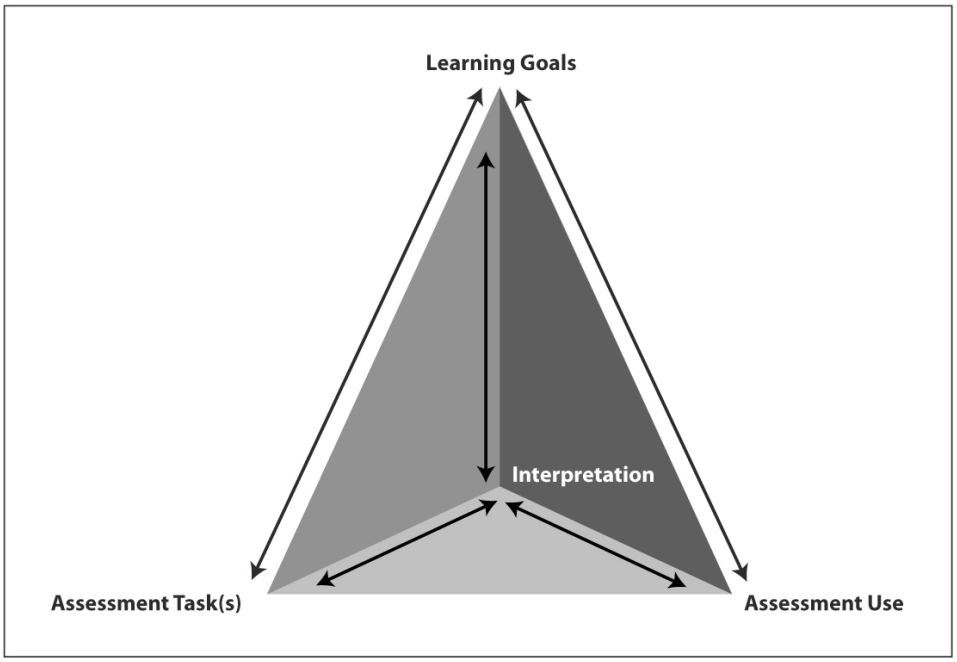

Figure 1. Figure 1 is adapted from the CAESL assessment model, developed in collaboration by CAESL colleagues. (See Herman et al., 2005; DiRanna et al., 2008).

Like KSWK, the tetrahedron starts (and ends) with specified goals for student learning, optimally conceptualized in a fine-grained representation of how students' knowledge in a domain develops. The assessment vertex indicates that tasks used for formative assessment need to align with specified learning goals and assessment purposes, whereas the

interpretation vertex reinforces the idea that responses from assessment must be analyzed and aggregated in ways that yield valid inferences about students' learning relative to the specified goals, and in psychometric terms, be grounded in a sound measurement system. Technical quality resides relationships between all three vertices, e.g., in the relationship between learning goals and tools, the appropriateness of the tasks for intended purposes, the reliability and accuracy of the analysis, and evidence that inferences are justified.

We add to this view of assessment quality attention to the process of using assessment and the evidence it provides to enhance student learning. That process also starts with goals for student learning: Commonly in systematic development (c.f. Bloom, 1968; Marzano & Kendall, 1996; Popham & Baker, 1970; McTighe & Wiggins, 2004), teachers are asked to establish learning goals, assess student status relative to them, plan and enact instruction accordingly, then assess the results and recycle as needed to achieve desired results. The assessment not only provides important technical information on which to base and refine teaching and learning, but the process adds value through other leverage points as well. For example, the process of identifying appropriate assessments may encourage teachers to clarify their goals and communicate their expectations explicitly to students, a practice which research associates with increases in learning. Even if goals are not explicitly communicated, the assessments themselves can serve a signaling function by communicating to students what is important to know and be able to do (Herman, 2006; Kirst & Venezia, 2004; Stecher & Chun, 2001), and can provide scaffolding for and an occasion to practice desired knowledge and skills (Sadler, 1989; Shepard, 2005).

In contrast to large-scale assessments, where the administration of an assessment is tantamount to being provided scores from it, classroom formative assessments generally demand that teachers do their own scoring and analysis/interpretation of results. Specifically called out in Figure 1, and the prime focus of the current study, this interpretation step often is a novel one for teachers, bringing with it both challenges in assuring the reliability and validity of inferences and new understandings of student conceptions and learning processes (Gearhart et al., 2005). As Gearhart et al. document, the process requires that teachers think more deeply about their primary learning goals, struggle to construct and apply relevant criteria, and transform their thinking about assessment and instruction from a focus on whether students are right or wrong ("got it" or "don't") to understanding and furthering student conceptions.

Once administered, scored, and interpreted, results must be applied in ways that support subsequent learning, e.g. through feedback to students, by adapting subsequent instruction, by identifying and providing supplementary activities for students who need help, perhaps

even with subsequent probing to get more detail to guide subsequent teaching and learning. Research, for example, is near unanimous on the positive effects of feedback in student learning and particularly on the value of descriptive feedback that provides students information on the quality of their performance and how to improve it (Kluger & DeNisi, 1996).

As noted above, the current study concentrates on only one component of our model, the quality of teachers' interpretation of assessment results, and how such quality, as defined by the accuracy of teachers' judgments, may be related to student performance. Our principal study questions thus are:

- How accurate are teachers' judgments of student learning, in the context of small study of middle school science?
- How does accuracy of teachers' judgments relate to student performance?

Methodology

The study involved seven experienced middle school science teachers from districts across California in the implementation of a unit from the Foundational Approaches in Science Teaching (FAST) curriculum developed by the University of Hawaii Curriculum Research and Development Group (Pottenger & Young, 1992). The unit included specially developed formative assessments at key juncture points, called "Reflective Lessons" (Ayala et al., in press). Study measures operationalized components of our model and included student performance on the reflective lessons, specially developed pre- and post assessments and multi-method data on teachers' interpretation and use of data from the reflective lessons. Analyses examined trajectories of student learning across time, within and across classrooms and the relationships between assessment practices and performance. Below is a summary of study methodology; additional details about instrumentation can be found in Herman et al., (2005).

Background on FAST

The study was based in FAST's introductory Physical Science unit on buoyancy, which we called *Why Things Sink and Float* (WTSF). With efficacy established by prior studies (Pauls, Young, Donald, & Lapitková, 1999; Pottenger & Young, 1992; Tamir & Yamamoto, 1977), the unit engages students through 12 investigations that guide students to progressively more sophisticated understanding of buoyancy by sequentially addressing the concepts of mass, volume, density, relative density, and their relationships to buoyancy. During these investigations, students apply a variety of science skills—observation,

prediction, summarizing results, and providing explanations—through a combination of individual, small group, and whole class activities involving discussion and reflection.

Background on Reflective Lessons

The formative assessments used in this study—called reflective lessons—were specially developed by Richard J. Shavelson and his colleagues at Stanford Education Assessment Laboratory (2005) to capture conceptual change at critical junctures in the unit's learning sequence, points at which students are expected to understand progressively the role of mass, volume, both mass and volume, density, and relative density in understanding why things sink and float (buoyancy). The reflective lessons were intended to provide teachers and students with important information about whether students were ready to move ahead with the curriculum or would benefit from additional work and instruction.

Each reflective lesson asked students individually and in writing to (a) interpret and evaluate a graph, (b) predict-observe-explain an event related to sinking and or floating, (c) answer a short essay question, and (d) predict-observe an unexpected event related to sinking and floating. Classroom discussion subsequent to each individual assessment was conceived to provide additional opportunities for teachers to elicit, probe, provide feedback on, and deepen students' understanding of unit concepts.

Reflective lessons (RL) were embedded after Investigations 4, 7, and 10 in the curriculum and are called RL4, RL7, and RL10 in the sections that follow. Guidance materials accompanying the assessments cued teachers on the use of the scoring rubrics and alerted them to common misconceptions and misunderstandings that commonly occurred at each assessment point. The materials also included suggestions for additional instructional activities that could benefit specific individual or groups of students who evidenced particular misconceptions and or were otherwise struggling. In this way, the materials provided support for teachers' interpretations and use of assessment results. (See Ayala et al., in press, for additional details of rationale and development of the reflective lessons and their supporting materials.)

The reflective lessons were scored using a progress variable (Wilson & Sloane, 2000) that reflected FAST's implicit developmental model for fostering student understanding of buoyancy (see Table 1). Students' responses were scored at one of eight levels, based on the quality of understanding their responses expressed. The scoring rubric operationally defined the nature of students' conceptual understanding at each level and those that were needed for a student to move to the next level of conceptual understanding. (See Kennedy, Brown, Draney, & Wilson, 2005, for additional detail about rubric and scoring).

Table 1

Buoyancy: WTSF Progress Guide (from Kennedy et al. 2005)

Level	What the student already knows		What the student needs to learn
RD	Relative Density Student knows that floating depends on having less density than the medium. • "An object floats when its density is less than the density of the medium."		
D	Density Student knows that floating depends on having a small density. • "An object floats when its density is small."		To progress to the next level, student needs to recognize that the medium plays an equally important role in determining if an object will sink or float.
MV	Mass and Volume Student knows that floating depends on having a small mass and a large volume. • "An object floats when its mass is small and its volume is large."		To progress to the next level, student needs to understand the concept of density as a way of combining mass and volume into a single property.
M V	Mass Student knows that floating depends on having a small mass. • "An object floats when its mass is small."	Volume Student knows that floating depends on having a large volume. • "An object floats when its volume is large."	To progress to the next level, student needs to recognize that changing EITHER mass OR volume will affect whether an object sinks or floats.
PM	Productive Misconception Student thinks that floating depends on having a small size, heft, or amount, or that it depends on being made out of a particular material. • "An object floats when it is small."		To progress to the next level, student needs to refine their ideas into equivalent statements about mass, volume, or density. For example, a small object has a small mass.
UF	Unconventional Feature Student thinks that floating depends on being flat, hollow, filled with air, or having holes. • "An object floats when it has air inside it."		To progress to the next level, student needs to refine their ideas into equivalent statements about size or heft. For example, a hollow object has a small heft.
OT	Off Target Student does not attend to any property or feature to explain floating. • "I have no idea."		To progress to the next level, student needs to focus on some property or feature of the object in order to explain why it sinks or floats.
NR	No Response Student left the response blank.		To progress to the next level, student needs to respond to the question.
X	Unscorable Student gave a response, but it cannot be interpreted for scoring.		

Note. WTSF = Why things sink and float.

Study Sample

Thirteen middle school teachers volunteered for the original study and participated in its initial training; of these, only seven teachers completed the unit and all data requirements. These seven teachers are the focus of the current research, which represents more of a case study rather than a firm empirical base.

Sampled teachers participated in a weeklong summer institute to orient them to the curriculum and its associated reflective lessons (i.e., formative assessments). To promote consistent implementation and understanding of both the curriculum and its assessment tools, the training provided expert demonstrations of and opportunities to engage in sample lesson activities, implement reflective lessons, and apply the scoring rubrics to student work.

Table 2

Summary of Student Background Variables.
Mean class values[1]

	Mean	*SD*	Minimum	Maximum
Class size	27.63	9.41	9	40
Percentage of female students	40.63	14.90	11	57
Percentage of minority students[2]	38.98	31.45	5	90
Percentage of low SES students[3]	17.99	27.55	0	72
Percentage of students classified as ELL[4]	2.64	5.12	0	15
Percentage of students classified as FEP[5]	17.56	22.53	0	61
Percentage of special needs students[6]	2.91	5.52	0	14

[1] The means of average class composition for 7 sites.
[2] Non-White
[3] Low socioeconomic status (SES) indicated through participation in the free lunch program.
[4] ELL = English Language Learner classification.
[5] FEP = Fluent English Proficiency classification.
[6] Students classified as having physical and or mental disabilities.

Study teachers taught in different middle schools across the state, reflecting various community and student characteristics, from a private school serving a relatively affluent community, to urban sites serving economically disadvantaged students of color and English language learners (ELL), to suburban and rural sites. A summary of the student demographics in our sampled classrooms is shown in Table 2. The data show a wide range of class sizes among observed teachers, ranging from 9 to 40 students. Representation of low socioeconomic status (SES) and minority students ranged from 0% to 72% and from 5% to 90%, respectively. Few classrooms had students classified as limited English proficient

(LEP) or as special needs students. Typically, our sampled classroom took 10 to 12 weeks to complete the FAST unit.

Data on the background and experience of study teachers revealed a highly capable and accomplished group in terms of prior science education and teaching experience. All but one teacher had an undergraduate major in science or in science education, and half had masters' degrees in these subjects. They averaged 13 years of prior science teaching experience. In fact, teacher characteristics are a major limitation of the study, in that the teachers were volunteers and cannot be considered representative of typical middle school science teachers. All had prior experience teaching FAST.

Data Sources and Study Variables

Study data sources included teacher logs in which teachers were asked regularly to estimate the percentages of their students currently at each level of the WTSF learning trajectory and repeated measures of student learning. The consistency between teacher judgments and researcher scores on the learning measures formed the accuracy of assessment variable for the study, as described below.

Teacher logs. After each investigation and each reflective lesson, teachers were asked to complete a web-based survey form in which they indicated both the proportion of their students who scored at each level of the WTSF progress trajectory and the sources of evidence they had used in assigning these ratings. Teachers might indicate for example, that 25% of their students were performing at the M (mass level), and that 25% were performing at the V (volume level), etc.

Student learning data: Pre–post measures and reflective lessons. The study used specially designed pre- and posttests to assess students' understanding of buoyancy. The pre-test for the purpose of the study reported here was composed of nine multiple-choice with justification items, and the posttest was constituted by these same items plus four reflective lesson activities from RL4. In addition to administering these pre- and posttests, teachers submitted students' responses to the reflective lessons, RL4, RL7, and RL10.

All five assessments—the pre-test, posttest, and three reflective lessons—were scored using the WTSF progress variable described earlier. Four CAESL researchers who were knowledgeable in science, familiar with the curriculum, and trained in a series of moderation sessions to consistently apply the rating scale conducted scoring.

The study's psychometric approach featured a multidimensional Rasch-based item response model, and the multidimensional random coefficients multinomial mode (Adams,

Wilson & Wang, 1997). A linking study was conducted to calibrate the items across all five assessments to enable valid comparisons of student proficiency based on the different sets of items used at the different points in time. To establish the relative difficulties of each set of items, four linking tests were administered to classes of students who were not in the study but who were similar to the study's student population. Each linking test enabled the direct comparison of items from two different assessments and with the posttest (which essentially contained the pre-test plus RL4) enabled the cross comparison of all (pre-test, RL4, RL7, RL10, posttest).

Weighted likelihood estimates of student proficiencies were calculated at each time point using ConQuest software (Wu, Adams, Wilson, & Haldane, 2005). These estimates ranged from -3 to +3 and can be interpreted as representing the proficiency at which a student has a 50% probability of responding at a particular WTSF level or higher. -3 represents the lowest level of proficiency on the scale, i.e., off target, and +3 the highest level of understanding, at the level of relative density. For example, for RL4, a student at approximately -1 has a 50% probability of answering at the level of "unconventional feature." Separation reliabilities for the resulting scores show more consistent results for pre- and posttests (.89 and .88, respectively) than for the reflective lessons, which ranged from .66 to .68. (See Kennedy et al., 2005, for additional details on assessment development, scoring, and underlying measurement models.)

Accuracy of teachers' interpretation. The correspondence between teachers' estimates of the distribution of student understanding and those derived from CAESL researchers' scoring of the same reflective lessons were used to compose a variable representing the accuracy of teachers' interpretation. That is, because teachers were asked after each reflective lesson to complete logs estimating the distribution of the class at each level of the WTSF progress variables, we were able to compare estimates based on these two sources. We treated the centralized scoring as the true score, translated these individual scores into their corresponding WTSF levels, and then computed the percentage of students in each class scoring at each of the WTSK levels. We then computed the percentage agreement between these centralized scores and teachers' estimates. So, for example, if the centralized scoring showed that 50% of a class' students were at the level of understanding mass, and 25% were at the level of understanding volume, and while another 25% were still at the level of productive misconceptions, but the teachers' estimates showed 50% at the level of mass and 50% at the level of volume, we would assign this teacher's accuracy score at RL4 as 75%. Admittedly, the comparison is imperfect in that teachers were asked only to provide gross estimates, whereas the centralized scores by their nature were more precise.

Analysis Strategies

We first computed descriptive statistics on the data, followed by analyses of the relationships between the accuracy of teachers' judgments and student performance. The longitudinal and nested character of the data encouraged us to plot individual student observed growth trajectories within each teacher. In addition to this graphical representation of longitudinal data, we also computed descriptive statistics for each variable within each teacher and for each time point. As seen both in Figures 2 and 3, and Table 4, student growth does not seem to be linear over time, which guides us to examine gain between adjacent time points. Moreover, the character of the descriptive results, both in terms of evidence of non-linearity of student growth trajectories and that of inconsistencies in individual teachers' accuracy, led us to examine relationships between accuracy and performance through a variety of lenses and analysis strategies.

The general hypothesis we were testing was that teachers' accuracy in judging student performance is positively associated with subsequent student learning, in that knowing where students are should benefit the sensitivity of subsequent teaching and learning. With sound knowledge of students' learning status, a teacher can better plan appropriate instruction. To examine this hypothesis, we examined the correlation between each teacher's accuracy at each time point and the subsequent learning gains for that teacher's students. In other words, for each teacher there were four measures of accuracy: at the pre-test, at RL4, at RL7, and at RL10; correspondingly there were four measures of student gain (from the pre-test to RL4, from RL4 to RL7, from RL7 to RL10, and from RL10 to the posttest). We term the correlation between these scores as the within-teacher accuracy and gain relationship.

Secondly, we examined how strongly teachers' overall accuracy is correlated with their average gain. Here, for each teacher, we averaged the four accuracy measures and the average of the four gains and looked at the correlation between the two. We termed this analysis the between-teacher accuracy and gain relationship.

Finally, the between-teacher accuracy and gain relationship at each time segment was examined using a 3-level hierarchical model (HM). This model basically investigated the extent to which differences in gain across teachers is related to differences in teachers' accuracy at each time point. This approach can be considered a further delineation of the second approach (between-teacher accuracy and gain relationship), because the HM results present the between-teacher accuracy and gain relationship at each assessment time point. Here, students' repeated measures are nested within students, who in turn are nested within teachers, and because of the non-linear pattern of growth across time periods, the model

essentially tests the relationship separately for each segment of the curriculum sequence or learning trajectory.

Results

Results of Student Learning Measures

Descriptive statistics for the student learning measures used in the study are displayed in Table 3. Recall that the scores are expressed as weighted likelihood estimates of student proficiencies, ranging from -3 to +3, that essentially represent the likelihood that a student would score at a particular level with 50% probability. The results suggest that initially, average classroom pre-test scores by classroom are more similar than one might expect based on the demographic differences in their student populations. Although all classrooms show increases from pre- to posttest, the patterns of growth across the time periods by teachers appear variable, and students on average end the unit with a level of understanding of buoyancy substantially below that expected by the curriculum. Specifically the goal of the curriculum is that students score at the relative density level of understanding by the end of the unit.

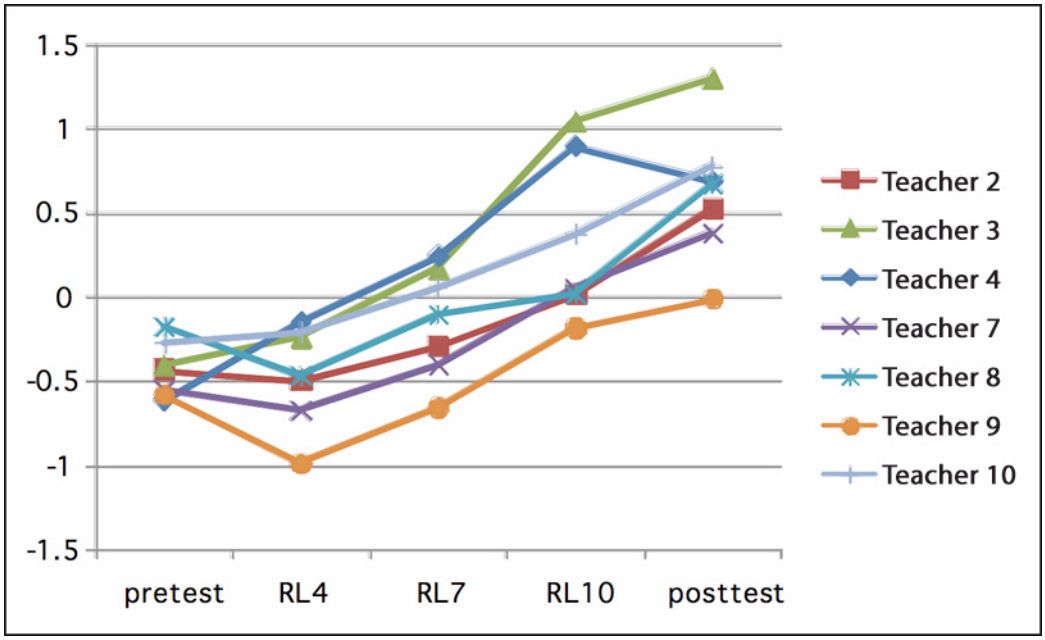

Figure 2. Mean growth trajectory for each teacher.

Apparent patterns are easier to see in Figure 2, which displays the mean growth trajectories for each of the seven teachers. Here again we see that students across all classes started at a similar level, but that scores tended to decrease slightly between the pre-test and

RL4. Although growth patterns are fairly similar across teachers, Teachers 3 and 4 stand out with the steepest levels of growth, even as Teacher 4 students' scores decline on the posttest.

Similarly, individual student growth trajectories over the time periods within each class/teacher also show substantial variability. Results for Teacher 3, as shown in Figure 3, are illustrative. They show individual students' scores bouncing around from one time point to the next. Note that several students, for example, who peak at RL10, but then fall back to lower levels of understanding at the time of the posttest.

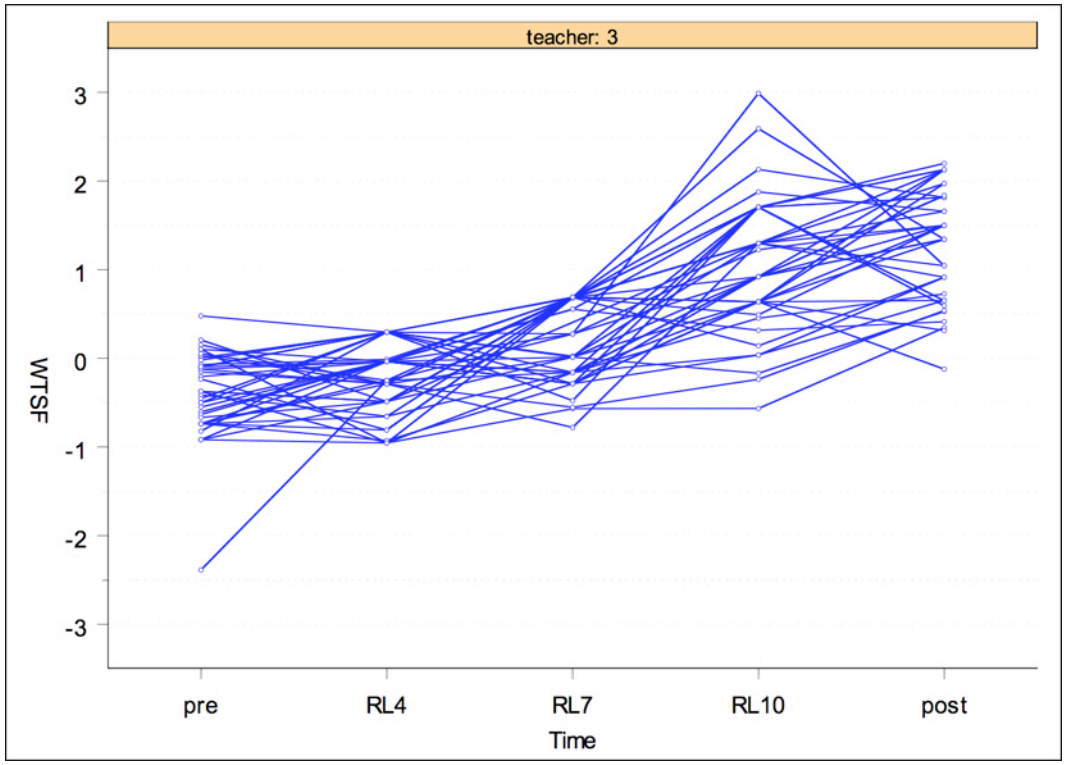

Figure 3. Individual growth trajectories within class.
Note. WTSF = Why things sink and float, RL = Reflective lesson.

Results for Accuracy of Teachers' Judgments

Table 3 shows the accuracy of teachers' judgments of students' level of understanding for each assessment occasion, expressed as the percentage match between student responses to each assessment as centrally scored by the research team and teacher estimates. A number of observations stand out: the relatively low agreement between teacher and researcher estimates (less than 50%), and the variability both across and within teachers. That is, some teachers on average are more accurate than others, but each teacher's accuracy varies considerably depending on the assessment time point, with no apparent pattern in the specific time points at which teachers are most or least accurate.

The average gains by teachers displayed in Table 3 reinforce the earlier observation: student scores tend to dip between the pre-test and the first reflective lesson, and gains tend to be highest near the midpoint of the unit.

Relationships Between Accuracy and Gains

As noted in the methodology section, we used multiple vantage points to examine the relationship between teacher accuracy and gains in student performance from multiple vantage points. Our analyses employed the four gain scores calculated from the learning data at each of five time points (pretest, RL4, RL7, RL10, and posttest) and the corresponding teacher accuracy data at each time point (see Tables 3 & 4). Our hypotheses addressed the relationship between the accuracy of teachers' judgments and their students' subsequent progress, under the assumption that more accurate judgments would lead subsequently to more effective teaching and learning for students. As a first lens, we calculated the coefficient between gains and accuracy measures for each teacher. For example, as shown in Table 5, the correlation coefficient for Teacher 2 between the four gains scores (RL4 to pretest; RL7 to RL4; RL10 to RL7; posttest to RL10) and the four accuracy measures at pretest, RL4, RL7, and RL10 is equal to .89. This coefficient captures the relationship between this individual teacher's accuracy and his or her classroom's subsequent average gain, which we termed the within-teacher accuracy and gain relationship. As can be seen in column 2 of Table 5, Teachers 2, 3, 4, and 7 show high positive relationships, i.e., .89, .93, .68, and .51, respectively. Interestingly, Teachers 3 and 4 showed the highest average gain among seven teachers. However, Teachers 8 and 10 show a small positive relationship, and Teacher 9 results reverses the pattern, by evidencing a strong negative relationship between accuracy and subsequent student gains. However, Teacher 9's context is unusual, in that the class contained only 9 students and these were composed exclusively of troubled youth. The average correlation across 6 teachers, excluding Teacher 9 is .57, which indicates a very sizable relationship.

Table 3

Descriptive Statistics of Outcome Measure at Each Time Point: Mean, *N*, *SD*

Teacher	Pretest (*N*, *SD*)	RL4 (*N*, *SD*)	RL7 (*N*, *SD*)	RL10 (*N*, *SD*)	Posttest (*N*, *SD*)
2	-0.424 (28, .610)	-0.489 (28, .416)	-0.282 (25, .362)	0.018 (26, .286)	0.533 (26, .633)
3	-0.398 (40, .485)	-0.234 (40, .395)	0.176 (39, .466)	1.049 (40, .831)	1.305 (40, .631)
4	-0.610 (32, .548)	-0.145 (32, .505)	0.245 (32, .524)	0.897 (32, .413)	0.687 (31, .492)
7	-0.543 (27, .749)	-0.661 (29, .342)	-0.392 (29, .439)	0.057 (29, .296)	0.385 (29, .365)
8	-0.171 (25, .820)	-0.460 (26, .693)	-0.099 (27, .729)	0.021 (23, .394)	0.676 (24, .503)
9	-0.570 (9, .558)	-0.975 (9, .478)	-0.648 (8, .543)	-0.179 (7, .517)	-0.004 (7, .264)
10	-0.262 (21, .642)	-0.197 (21, .548)	0.063 (21, .949)	0.381 (21, .677)	0.784 (21, .674)
Avg.	-0.425	-0.452	-0.134	0.320	0.624

Note. RL = Reflective lesson.

Table 4

Descriptive Statistics of Classroom Average Gain and Teacher's Accuracy of Student Rating

	Average gain				% Accuracy			
Teacher	RL4–pre	RL7–RL4	RL10–RL7	Post–RL10	RL4	RL7	RL10	AV
2	-0.065	0.189	0.489	0.515	52	62	54	40.4
3	0.164	0.404	1.283	0.256	48	64	35	49.8
4	0.465	0.390	1.042	-0.211	16	62	75	45.0
7	-0.143	0.269	0.718	0.328	31	43	18	31.4
8	-0.340	0.373	0.550	0.522	17	48	39	33.0
9	-0.405	0.337	0.720	0.026	22	31	36	35.8
10	0.066	0.259	0.577	0.403	37	42	58	51.0
Avg.	-0.037	0.317	0.768	0.263	31.9	50.3	45.0	40.9

Table 5

Correlation Between Teacher's Accuracy of Student Rating and His or Her Classroom Average Gain

Teacher	Corr (gain, accuracy)	Avg. accuracy	Avg. gain
2	0.89	40.4	0.282
3	0.93	49.8	0.527
4	0.68	45.0	0.421
7	0.51	31.4	0.293
8	0.17	33.0	0.276
9	-0.77	35.8	0.170
10	0.22	51.0	0.326
Avg.	0.38	40.9	0.33

Note. Within-teacher Avg. Corr = 0.38, between-teacher Corr = 0.68

Secondly, we examined how strongly teachers' average accuracy is correlated with their average gain across all time periods, as shown in Table 5. In contrast to the within-teacher accuracy and gain relationship, we term this coefficient the between-teacher accuracy gain relationship. As can be seen in the third and fourth column in Table 5, there is a tendency for teachers with higher average accuracy to have higher gains in student scores than those with lower average accuracy. The resulting correlation coefficient is .68, which confirms the between-teacher relationship.

Thirdly, the between-teacher accuracy and gain relationship at each time segment was examined using a 3-level hierarchical model (HM) showing the extent to which differences in gain across teachers is related to differences in accuracy across teachers at each time segment. Table 6 presents the results from a 3-level HM where students' repeated measures are nested within students who are in turn nested within teachers.

Table 6

Result Based on 3-Level Hierarchical Model: Relationship Between Gain and Teacher Accuracy

Fixed effect:	Estimate	*SE*	*t*-value	*df*	*p*-value
Status at pretest	-0.423	0.047	-9.00	185	0.000
Model for gain 1 (RL4–pretest)					
Avg. gain	-0.007	0.116	-0.06	5	0.957
Accuracy at pretest	0.025	0.003	7.80	5	0.000
Model for gain 2 (RL7–RL4)					
Avg. gain	0.321	0.048	6.70	895	0.000
Accuracy at RL4	0.005	0.002	2.01	895	0.045
Model for gain 3 (RL10–RL7)					
Avg. gain	0.473	0.117	4.03	5	0.014
Accuracy at RL7	0.020	0.003	5.98	5	0.000
Model for gain 4 (posttest–RL10)					
Avg. gain	0.250	0.098	2.55	5	0.050
Accuracy at RL10	-0.010	0.003	-3.88	5	0.016
Variance components:	Estimate	*df*	Chi-square	*p*-value	
Level-1	0.209				
Level-2					
Status at pretest	0.196	181	353.8	0.000	
Gain (RL4–pretest)	0.079	175	231.6	0.003	
Level–3					
Gain 1 (RL4–pretest)	0.074	5	60.14	0.000	
Gain 3 (RL10–RL7)	0.078	5	34.48	0.000	
Gain 4 (posttest–RL10)	0.050	5	21.45	0.001	

Note. RL = Reflective lesson.

First, the estimate of status at pretest is equal to -0.423, which corresponds to a level of understanding between "productive misconception" and "mass." As seen in the estimates of teacher accuracy effects on the gains, teachers accuracy has statistically significant effects on student gains. For example, the estimate of teacher accuracy at pretest shows a 0.025 effect on the first gain between RL4 and pretest, with a *p*-value 0.00. Although this indicates that a one percentage increase of teacher accuracy translates into only an increase of 0.025 point in the outcome score, there would be approximately half standard deviation point increase (i.e., 0.25) if the accuracy increases by 10%. Likewise, the estimate of the relationship between gain 2 (RL7–RL4) and teacher accuracy at RL4 is 0.005, which is also statistically

significant, as is the case for the estimate of the relationship between gain 3 (RL10–RL7) and teacher accuracy at RL10. Note, however, that the gain between the posttest and RL10 is negatively associated with teacher accuracy at RL10. This is partially because Teacher 4 distorts the relationship. The accuracy measure for Teacher 4 is the highest one among seven teachers, but the average gain for students in that class is negative for the final segment (see Table 4).

Discussion and Implications

This study has examined the accuracy of teachers' judgments of students' understanding and the relationship of such accuracy to middle school students' learning in the context of implementing the FAST Physical Science Curriculum on Why Things Sink and Float. The study is based on a framework for sound formative assessment practice that combines attention to both the quality and validity of assessment evidence and the quality of the use of such evidence to guide instruction and provide feedback to students. By concentrating on the accuracy of teachers' judgments of students' learning drawn from quality curriculum-embedded assessments, the study explores the contribution of one component of the framework. The sample size is admittedly very small, providing more of a case study than a firm empirical base, severely constraining any generalization of findings. Nonetheless, results are promising in terms of highlighting both the importance of and some potential challenges in assuring quality formative assessment practice.

First, the study lends support for the power of assessment in improving student learning. Assessment enables teachers to know where their students are performing relative to learning goals, and despite a very small sample and clearly imperfectly reliable measures, study results show that the more accurate teachers are in their knowledge of where students are, the more effective they may be in promoting subsequent subject learning. This was true when we looked within teachers over the course of the FAST unit, between teachers across the unit, and in more detail between teachers at each assessment time point. That is, in examining each teacher's accuracy at each of four time points during the unit in relationship to the subsequent progress of his or her students, we found generally positive relationships. Similarly, when we looked across the unit and looked at the correlation between each teacher's average accuracy and the average of their students' learning gains, we found a strong positive relationship. Finally, when we used HM methodology to look at the relationship between teachers' relative accuracy at each time point and their students' subsequent performance, we also found significant positive effects. Although the practical size of the effects may be small, results are striking in their consistency.

We believe these results also reinforce the potential value of our underlying framework in demonstrating the importance of one of its components. That is, quality of interpretation does seem to matter. Quality or validity in assessment requires that results be accurate for intended decision-making purposes. In this study we started with assessment tasks and scoring rubrics that had been carefully and professionally designed to be aligned with learning goals and to provide detailed information from which to gauge student progress and take subsequent action (courtesy of Ayala et al., in press; Kennedy et al., 2005); yet the validity of such assessments ultimately rests on whether teachers can appropriately interpret and use the results. In underscoring the importance of the framework's interpretation vertex, study results also point to potential challenges in assuring the accuracy of teachers' interpretations. As judged by the consistency between teachers' judgments about the distribution of their students' levels of understanding at each assessment point with that obtained by researcher scores on the same assessments, results showed considerable room for improvement. Certainly there are threats to the validity of our accuracy metric: teachers may or may not have looked carefully at each student's responses prior to their estimates, they may not have taken the estimates seriously or even completed the estimates on a timely basis; some differentiation between the two sources would be expected based on the different metrics on which the two measures were initially based (teachers were asked for gross estimates whereas the researcher estimates were based on individual scores); and there are some flaws in assuming that the scores arrived at by the central research team are "true." In addition, this was the first time that teachers had used these assessments and the scoring rubric and one might expect them to be more accurate with subsequent iterations. Nonetheless, that accuracy was highly related to student performance and that overall average accuracy was less than 50% gives pause, as does the variation in accuracy within teachers.

At the same time, the patterns of individual student growth trajectories within each class demonstrate the difficulty in getting a firm fix on individual student understanding. Although within classes students on average showed consistent upward (if modest) trajectories, the assessed levels of understanding for some students were quite erratic from one assessment occasion to the next, and there were many examples of students losing significant ground, i.e., by subsequently scoring at lower levels on the scoring guide. How valid are the results at each time point? Is the erratic performance a function of relatively weak reliability? Were students confused by the more advanced content dealt with in subsequent parts of the unit? Or is this the way learning naturally occurs, in fits and starts,

some up, some down? How seriously should we take individual student results at one point in time?

That accuracy in teachers' interpretation of learning cannot be assumed, even for a group of highly experienced and science-knowledgeable teachers who claim considerable capability in assessment, seems a clear implication of the study results. Also, issues of accuracy in interpretation may help to explain the absence of findings in prior studies of assessment use, ours among them (see, for example, Herman et al., 2005; Yin et al., in press). That is, accuracy in assessment seems a necessary precursor to the use of results benefiting student learning. With inaccurate interpretation, it is difficult to provide useful feedback or to optimize next steps for teaching and learning. Our findings suggest that studies of assessment use ought to take account of assessment accuracy.

In summary, we believe our study documents the value of a systematic approach to the study of formative assessment practice and to an understanding of whether and how it supports student learning. As we better uncover the conditions and precursors to good practice, we will be better able to foster the policies, capacities, and resources that can make such practice a reality. Until we do so, formative assessment risks remaining a slogan and achieving its potential value, elusive.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson & G. Engelhard (Eds.), *Objective measurement: Theory into practice* (Vol. 4, pp. 139–155). Norwood, NJ: Ablex Publishing.

Ayala, C., Shavelson, R. J., Brandon, P., Yen, Y., Furtak, E., Ruiz-Primo, et al., (in press). From formal embedded assessments to reflective lessons: The development of formative assessment suites. *Applied Measurement in Education*.

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

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

Black, P., & Wiliam, D. (2004). 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.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment, 1*(2), 1–8. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

DiRanna, K., Osmundson, E., Topps, J., Barakos, L., Gearheart, M., Cerwin, K., et al., (2008). *Assessment-centered teaching: A reflective practice.* Thousand Oaks, CA: Corwin Press.

Dorr-Bremme, D., & Herman, J. L. (1986). *Assessing student achievement: A profile of classroom practices*. In E. L. Baker (Ed.), (CSE Monograph Series in Evaluation, No. 11). Los Angeles: University of California, National Center for Research, on Evaulation, Standards, and Student Testing (CRESST).

Gearhart, M., Nagashima, S., Pfotenhauer, J., Clark, C., Schwab, C., Vendlinski, T., et al. (2005). Developing expertise with classroom assessment in K–12 Science: Learning to interpret student work. Interim findings from a two-year study. *Educational Assessment Journal.*

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

Hamilton, L., & Stecher, B. (2006, April). *Using test score data in classrooms.* Paper presented at the annual meeting of the American Educational Research Association (AERA), San Francisco, CA.

Heritage, M., & Yeagley, R. (2005). Data use and school improvement: Challenges and prospects. In J. L. Herman & E. L. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement*. *104th Yearbook of the National Society for the Study of Education (NSSE), Part 2*. Malden, MA: Blackwell Publishing.

Herman, J. L. (2006). Challenges in integrating standards and assessment with student learning. *Measurement: Interdisciplinary Research and Perspectives, 4*(1,2).

Herman, J. L., & Gribbons, B. (2001). *Lessons learned in using data to support school inquiry and continuous improvement: Final report to the Stuart Foundation.* (CRESST Tech. Rep. No. 535). Los Angeles: University of California, National Center for Research, on Evaulation, Standards, and Student Testing (CRESST).

Herman, J. L., & Heritage, M. (2007, June). *Moving from piecemeal to effective formative assessment practice: Moving pictures on the road to student learning.* Paper presented at the annual Council of Chief State School Officers (CCSSO) National Conference on Large Scale Assessment, Session 143, Nashville, TN.

Herman, J. L., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2005, April). The nature and impact of teachers' formative assessment practices. In J. L. Herman (Chair), *Building science assessment systems that serve accountability and student learning: The CAESL model.* Symposium conducted at the annual meeting of the American Educational Research Association, Montréal, Canada.

Kennedy, C., Brown, N., Draney, K., & Wilson, M. (2005, April). *Using progress variables and embedded assessment to improve teaching and learning.* Paper presented at the annual meeting of the American Education Research Association, Montréal, Canada. Retrieved February 29, 2008, from http://bearcenter.berkeley.edu/publications/pubs.php#Presentations

Kirst, M., & Venezia, A., (Eds.). (2004). *From high school to college: Improving opportunities for success in postsecondary education.* San Francisco: Jossey-Bass.

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*(2), 254–284.

Marzano, R., & Kendall, J. (1996). *Designing standards-based districts, schools, and classrooms.* Alexandria, VA: Association for Supervision and Curriculum Development.

McMorris, R., & Boothroyd, R. (1993). Tests that teachers built: An analysis of classroom tests in science and math. *Applied Measurement in Education, 6*(4), 321–342.

McTighe, J., & Wiggins, G. (2004). *Understanding by design professional development workbook.* Alexandria, VA: Association for Supervision and Curriculum Development.

National Research Council (NRC). (2001). 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., Donald, B., & Lapitková, V. (1999). Laboratory for learning. *The Science Teacher, 66*(1), 27–29.

Perie, M., Marion, S., & Gong, B. (2007). *A framework for considering interim assessments.* Dover, NH: National Center for the Improvement of Educational Assessment.

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.

Popham, W. J. (1987). The merits of measurement-drive instruction. *Phi Delta Kappan*, *68,* 679–682.

Popham, W. J., & Baker, E. L. (1970). *Systematic instruction*. Englewood Cliffs, NJ: Prentice-Hall.

Pottenger, F. M., & Young, D. B. (1992). *The local environment: FAST 1 foundational approaches in science teaching.* Honolulu, HI: University of Hawaii, Curriculum Research and Development Group.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford, & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction*. Boston: Kluwer Academic Publishers.

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

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

Shavelson, R. J. (2006). *On the integration of formative assessment in teaching and learning with implications for teacher education.* Paper prepared for the Stanford Educational Assessment Laboratory and the University of Hawaii Curriculum Research and Development Group. Retrieved October 19, 2006, from http://www.stanford.edu/dept/SUSE/SEAL/Reports_Papers/Paper.htm

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

Shepard, L. A. (2005). Linking formative assessment to scaffolding. *Educational Leadership 63*(3), 66–71.

Stecher, B. M., & Chun, T. (2001, April). Schools in transition: A statewide profile of the impact of Washington's educational reform. Paper presented as part of the symposium, *A reform in the making: Early effects of a high-stakes, standards-based state reform on schools and classrooms,* at the annual meeting of the American Educational Research Association, Seattle, WA.

Stiggins, R. J. (2005). *Student-involved assessment FOR learning*. (4th ed.). Upper Saddle River, NJ: Pearson Merrill Prentice Hall.

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.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. In G. M. Whipple (Ed.), *The measurement of educational products. Seventeenth yearbook of the National Society for the Study of Education, Part II* (pp. 16–24). Bloomington, IL: Public School Publishing Company.

Tyler, R. W. (1948). How can we improve high-school teaching? *The School Review, 56*(7), 387–399.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*(2), 181–208.

Wu, M., Adams, R., & Wilson, M. (2005). *ACER ConQuest, version 1.0.* [Computer software] Melbourne, Australia: Australian Council for Educational Research.

Yin, Y., Ayala, C., Shavelson, R. J., Ruiz-Primo, M., Tomita, M., Furtak, E., et al., (in press). On the measurement and impact of formative assessment on students' motivation, achievement and conceptual change. *Applied Measurement in Education*.