CRESST REPORT 770

Joan L. Herman

Ellen Osmundson

David Silver

CAPTURING QUALITY IN FORMATIVE ASSESSMENT PRACTICE: MEASUREMENT CHALLENGES

JUNE, 2010

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

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

Capturing Quality in Formative Assessment Practice: Measurement Challenges

CRESST Report 770

Joan L. Herman, Ellen Osmundson, & David Silver
CRESST/University of California, Los Angeles

June, 2010

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 Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported by WestEd, grant #5387 S07-091, as administered by the Institute of Education Sciences (IES), U.S. Department of Education (Education Research Grants in Mathematics and Science—*Diagnostic Embedded Classroom Assessment—An Efficacy Study*) and the Lawrence Hall of Science, University of California at Berkeley.

The findings and opinions expressed herein are those of the author(s) and do not necessarily reflect the positions or policies of WestEd, Lawrence Hall of Science, or the US Department of Education (IES).

To cite from this report, please use the following as your APA reference: Herman, J., L., Osmundson, E., & Silver, D. (2010). *Capturing quality in formative assessment practice: Measurement challenges.* (CRESST Report 770). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

CAPTURING QUALITY IN FORMATIVE ASSESSMENT PRACTICE: MEASUREMENT CHALLENGES

Joan L. Herman, Ellen Osmundson, and David Silver
CRESST/University of California, Los Angeles

Abstract

This study examines measures of formative assessment practice using data from a study of the implementation and effects of adding curriculum embedded measures to a hands-on science program for upper elementary school students. The authors present a unifying conception for measuring critical elements of formative assessment practice, illustrate common measures for doing so, and investigate the relationships among and between scores on these measures. Findings raise important issues with regard to both the challenge of obtaining valid measures of teachers' assessment practice and the uneven quality nature of current teacher practice.

Introduction

Using evidence from studies of feedback, mastery learning, special education, and other specific teaching practices, Black and Wiliam (1998), concluded that formative assessment is a powerful classroom intervention, particularly for low achieving students (OECD, 2005). Heeding their advice and spurred on by researchers from diverse theoretical perspectives (See reviews by Shepard, 2005; Herman, 2010; James et al., 2007) and those from practitioner communities, policymakers across the world are considering formative assessment as a primary approach to educational reform (OECD, 2005; CCSSO, 2008). Current federal policy in the United States is a case in point: Billions of dollars are being invested in Race to the Top initiatives that put new standards and assessments front and center: amongst other things, results from new assessments are intended to populate state databases to inform improvement and to fuel efforts to turn around struggling schools. In addition, the federal government is investing $350 million in assessment development grants aimed at the development of state standards-based assessment systems that better support teaching and learning. While system development focuses on testing for purposes of accountability, for the first time the federal program includes attention to formative assessment and the development of local capacity and local assessment resources.

These indeed are promising developments for pushing formative assessment to fruition in classroom practice. They acknowledge and work toward remedying the need for classroom tools to assess and support student learning. Yet at the same time, recent studies reveal challenges in implementing quality formative assessment and show non-robust results with regard to effects on student learning (Herman, Osmundson, Ayala, Schneider, & Timms, 2006; Furtak, et al., 2008). Just as the concept of formative assessment itself underscores the central role of evidence—learning data—in an effective teaching and learning process, so too do policymakers and practitioners need evidence on which to build effective formative practices. Toward this latter goal, this report explores measures of formative assessment practice using data from a study of the implementation and effects of adding curriculum embedded measures to a hands-on science program for upper elementary school students. In the sections that follow, we present a unifying conception for measuring critical elements of formative assessment practice, illustrate common measures for doing so, and investigate the relationships among and between scores on these measures. Findings raise important validity issues and critical concerns in assessing quality practice.

Perspective: Formative Assessment Construct

Synthesizing distinct views of formative assessment, our core conception is grounded in modern validity theory about the meaning of quality assessment *measures* (AERA, APA, NCME, 1999; NRC, 2001) and adds to it concerns for the quality of the assessment *process*. Echoing the *Knowing What Students Know* assessment triangle, our conception of the validity of formative assessment measures rests with connections among and between the learning construct(s) being measured, the task(s) designed to elicit student responses, and *interpretive* frameworks used to make sense of and act on student responses. The validity argument, in part, rests on evidence that teachers' formative strategies elicit evidence of learning related to the goals, at the *level of detail* needed, and yield appropriate inferences for subsequent instructional decision making. Similarly, a quality formative assessment *process* also starts with specified (and significant) learning goals, iterative process of assessment, interpretation, and use of evidence to guide subsequent teaching and learning, to reduce the gap between students' current understandings and expected learning goals (See, for example, Heritage, 2010; Bell & Cowie, 2001). Figure 1 displays these key elements and processes, which we view as essential constructs in measuring/assessing the quality of formative practice. Effective practice requires the intertwining of quality evidence and a quality process or use: one without the other is counter-productive.

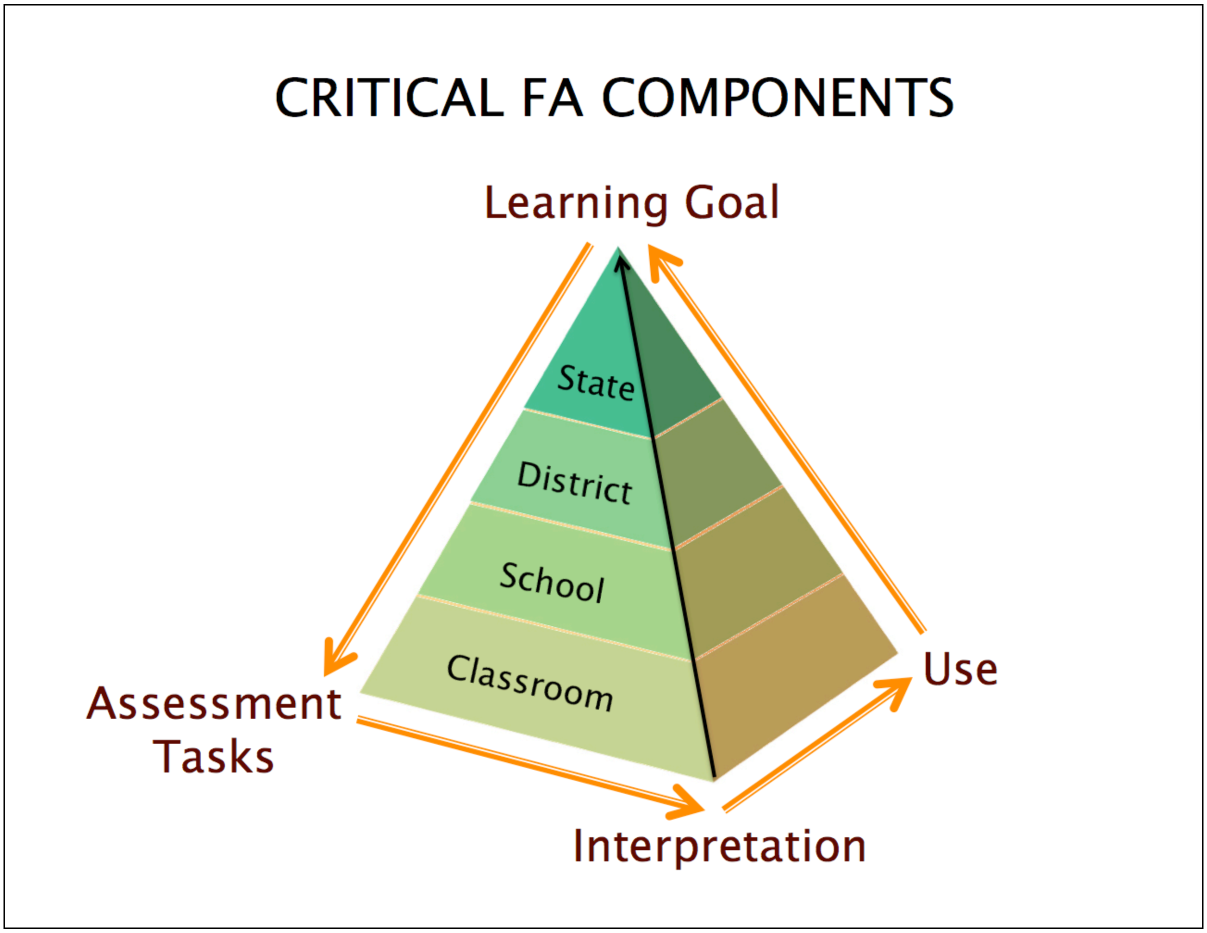


Figure 1. Critical Formative Assessment Components.

Illustrative Study: Design and Available Measures

A randomized field study of the effects of incorporating new formative assessments into an upper elementary hands-on science curriculum program provides the context for examining measures of these formative assessment constructs, and highlights the importance of differentiating the quality of the assessment from the process of assessment. The initial phase of the study involved 40 teachers, who were randomly assigned by the school to treatment (revised program with curriculum-embedded assessments) or control (traditional curriculum) conditions. This initial phase involved teachers in professional development designed to deepen content knowledge, and for treatment teachers, sessions to support the analysis and interpretation of student work, as well as next steps in instruction. Each group engaged in a practice year of implementing the curriculum with fidelity in preparation for the Year 2 test of treatment impact. Given the focus of the treatment, the study is using a variety of methods to collect data on teachers' assessment practices, including teacher surveys, logs, and direct measures of teachers content-pedagogical knowledge, including teachers' ability to interpret and act upon student work. Measures of student learning, along with the teacher measures discussed below, will be collected for the second phase of the study.

Sample

Schools and teachers from one southwestern state who had prior experience teaching the targeted science curriculum program were recruited for the study. Table 1 shows the demographics of the 39 teachers for whom data was available. For the purposes reported here, there is no reason to differentiate treatment or control status. The data in Table 1 show that both groups were very similar demographically—teachers were predominantly Caucasian females, were experienced teachers as well as experienced with the target curriculum materials, and had engaged in substantial professional development in science in the 2 years prior to the study (See Table 1).

Table 1

Cohort 1 Pilot Year (2008–2009): Teacher Demographic Information

Descriptor	Control *N* = 19	Treatment *N* = 20
Sex		
Male	1	0
Female	18	20
Ethnicity		
White	17	17
Hispanic/Latino(a)	2	2
Native American/African American	0	1
Other	0	0
Highest Degree Received		
Bachelor's + Credential	5	6
Bachelor's + Credential + Units Beyond	3	4
Master's	3	5
Master's + Units Beyond	8	5
Teaching Credential*		
General Elementary	18	17
General Secondary	1	1
Special Emergency	2	3
Multiple subject	1	1
Single subject	2	2
Bilingual	4	6
Administrative	1	1
Other: (Early childhood, TESOL, guidance, special education, science endorsement)	4	5
Years of experience teaching elementary grades		
Average number	12 years	8.4 years
Range of years teaching	1–32 years	2–25 years
Years teaching science curriculum unit		
Average number of years	3 years	2.6 years
Range of years teaching	1–11 years	2–12 years
Number of science Professional Development hours in the past 2 years		
Average number of hours	19.6 hours	21.3 hours
Range of hours	4–100 hours	2–80 hours

Note. *Teachers may hold multiple credentials.

Study Instrumentation

Table 2 summarizes the data sources for the current report. A direct measure of teachers' content pedagogical knowledge provides a window for examining the *quality* of teachers' assessments, while teacher surveys and weekly logs provide self-report data on the *assessment processes* in which teachers were engaged. Observation and interview data were collected for only a small subsample (see Table 2).

Table 2

Summary of Data Sources

Measure type	Understand content goals/ Depth of knowledge	Establish/ Assess goals	Assessment interpretation	Assessment use
Direct measure	X		X	X
Self report				
Survey	X	X	X	X
Weekly log				X
Interview		X	X	X
Observation		X		X

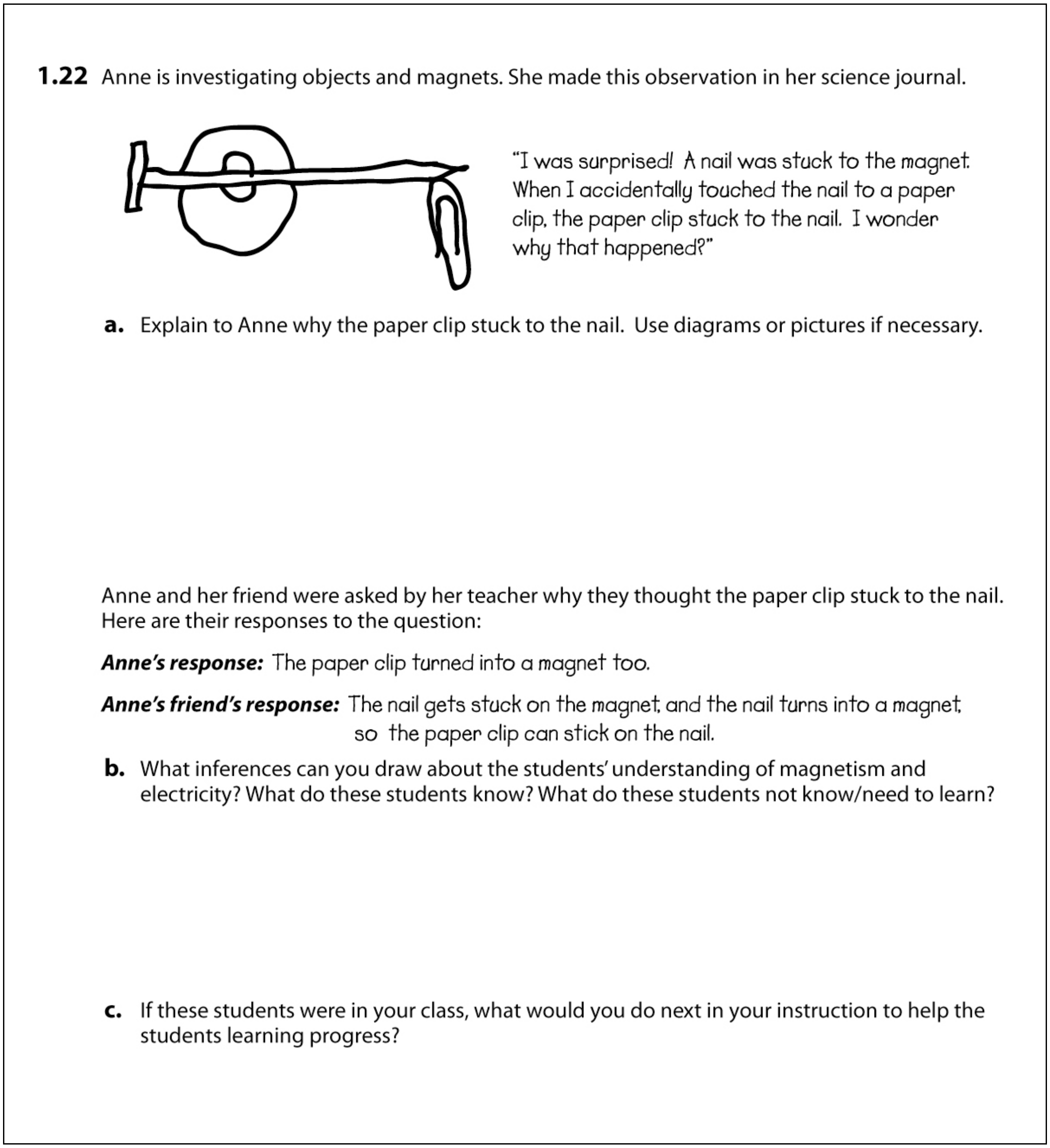
Direct measure of teacher content-pedagogical knowledge. A specially developed teacher-content-pedagogical-knowledge measure focused on magnetism and electricity, the topic for one of two curriculum units all participants implemented for the full study. The measure was administered before the start of the study and will be readministered at its end, after teachers have implemented the curriculum twice, in 2 subsequent years. Three item types corresponded to the different aspects of assessment quality delineated above: (a) content knowledge as a proxy for teachers' ability to understand learning goals; (b) items that asked teachers to analyze students' work as a proxy for quality of interpretation; and (c) items that queried teachers' ability to formulate next instructional steps, which corresponded to quality of use. All item types, items, and scoring schemes were reviewed for scientific accuracy and alignment with the subject curriculum and revised, when necessary, by two science education experts. Task formats were adapted from those used by other researchers to investigate teacher knowledge (Heller, Daehler, Shinohara, & Kaskowitz, 2004; Heritage, Kim, Vendlinski, & Herman, 2009; Hill, Schilling, & Ball, 2004).

The content knowledge measure comprised the first section of the assessment and contained 29 multiple-choice and short explanation items originally intended for students.

These items were culled from the subject curriculum's Magnetism and Electricity module, as well as from publicly released assessments: NAEP and TIMSS 4th-grade assessments on magnetism and electricity (Olson, Martin, & Mullis, 2008). Consistent with the subject curriculum's coverage, three topics were assessed: (a) magnetism, (b) electricity, and (c) electromagnetism.

The reliability of this measure was disappointing: subscales by sub topic (magnetism, electricity, and electromagnetism) achieved alphas of .44–.46, and even after deleting poor performing items, the highest performing scale reached .65, as did the combined set of items. As the items constituting the measure were designed for students, there were problems of range restriction and limited variation in performance.

Analysis and interpretation tasks constituted Section 2 of the assessment. These tasks were structured as follows: first, teachers were presented with an explanation task for students, one of the same ones they answered in Section 1 above. Then teachers were provided with student responses to the question and asked to analyze the students' responses to draw inferences about student understandings—what students understood, and what alternative conceptions might be evident. A third part of the section asked teachers to prescribe next instructional steps for the student(s), the indicator of quality of use, as described further. Figure 2 shows a sample item from the Magnetism and Electricity Module that follows the sequence described.



1.22 Anne is investigating objects and magnets. She made this observation in her science journal.

"I was surprised! A nail was stuck to the magnet. When I accidentally touched the nail to a paper clip, the paper clip stuck to the nail. I wonder why that happened?"

a. Explain to Anne why the paper clip stuck to the nail. Use diagrams or pictures if necessary.

Anne and her friend were asked by her teacher why they thought the paper clip stuck to the nail. Here are their responses to the question:

Anne's response: The paper clip turned into a magnet too.

Anne's friend's response: The nail gets stuck on the magnet, and the nail turns into a magnet, so the paper clip can stick on the nail.

b. What inferences can you draw about the students' understanding of magnetism and electricity? What do these students know? What do these students not know/need to learn?

c. If these students were in your class, what would you do next in your instruction to help the students learning progress?

Figure 2. Teacher Content Survey: Magnetism and Electricity Module (The Regents of the University of California, 2005).

Scores for this portion of the content survey (Figure 2, Part B) were based on a 3-point scale, derived from expert ratings of teacher responses. (See Table 3 for the summary of coding guide). All six analysis and interpretation items were double scored by two researchers, and inter-rater reliability (exact match) was calculated at 88%; differences in scores were discussed and resolved. Internal consistency of this item set was weak, at 0.54–0.65, depending on whether responses to two problematic items were included.

Table 3

Teacher Analysis and Interpretation of Student Work Coding Guide

Score	Description
3	Complete response, scientifically accurate identification of student level of understanding. E.g., "students understand that metal conducts electricity, but don't understand that all metals don't stick to magnets."
2	Partial response, mostly scientifically accurate identification of student level of understanding. E.g., "students understand magnetism and how it works."
1	Minimal response, minimal level of accuracy identification of student level of understanding. E.g., "student's understanding is incorrect. Ss need to go back and retest items in magnetism chart."
0	No response or response that indicates teacher does not understand the student response. E.g., "I'm not sure what the student is thinking."

Use: Next Steps for Instruction Tasks, described above, were scored on a 3-point scale, as devised by content experts and then revised based on teacher responses (See Table 4). Responses to each of the six tasks comprising this portion of the assessment were doubled scored by two researchers. The percentage of exact agreement, 68%, was lower than that of the analysis tasks, but differences were discussed and resolved to arrive at final consensus scores. Despite the challenge of rater agreement, internal consistency for this item set was higher than that achieved for the Analysis and Interpretation items. Including all six Next Step Tasks, reliability was computed at .74, and with two outliers excluded, .82.

Responses to both Analysis and Interpretation and Next Step Tasks were combined to achieve a more reliable scale. Alpha for the combined set reached .81.

Table 4

Teacher Instructional Next Steps Coding Scheme

Score	Description
3	Detailed, content-specific next instructional steps indicated. Response takes into consideration students' current level of understanding. E.g., "Next we'll investigate induced magnetism with different objects. Discuss why some worked and others did not. Record findings."
2	General, content-general instructional next steps indicated. Response alludes to "general" level of student understanding. E.g., "Next, I need to help Ss deepen understanding of electromagnets by showing different models."
1	Broad, vague instructional next steps indicated. Response does not take into consideration students' level of understanding. E.g., "Next, students need more experience with magnets, review content, more practice, journal our experiences."
0	No response or response that indicates teacher does not understand the student response. E.g., "I'm not sure what I would do next in instruction."

Teacher self report of content-pedagogical knowledge. Complementing the direct measures, we also asked teachers to rate their knowledge. On the pre-intervention survey teachers were asked to rate how well qualified they felt to teach fourth grade students a range of science topics, using a 5-point scale, from 1 (*not qualified*) to 5 (*highly qualified*). For purposes here, we report only on items related to magnetism and electricity, reflecting key ideas in the subject curriculum unit:

- Magnetic forces,
- How electrical circuits are designed,
- How electricity produces magnetic effects, and
- Overall magnetism and electricity.

Responses to these items, plus one asking them how well they understood the curriculum unit's learning goals, were summarized in a single scale (alpha = .81).

Teacher self-report of assessment practices. Items from the pre-institute survey asked teachers about the frequency with which they engaged in various aspects of the assessment process: setting and communicating goals, aligning assessments with learning and instruction; analyzing/interpreting student work (individual, small, and large group discussion, and use of curriculum-based strategies to (discussion, response sheets, performance assessments, and notebooks) to assess student understanding. Table 5 describes the items comprising and reliability for each of these scales.

Table 5

Teachers Self-Reported Assessment Practices

Assessment Processes	Items	Alpha
Establish/communicate goals	3.3a–c	.79
Align assessment w/ goals	3.5a, 3.3d, f, 3.3e, g	.78
Analyze/interpret	3.5c, d, e, f	.78
Use assessment	2.6a–f	.70

Teacher logs: use of curriculum and assessment. Teacher Logs were designed to measure teachers' use and implementation of the science unit curriculum and assessments, and to provide a general gauge of fidelity of implementation for various program constructs and ideas. Teachers were assigned IDs and logins, and were asked to report their instructional and assessment activities on a weekly basis. General reporting categories in the teacher log included: (a) amount of time students engaged with the curriculum; (b) amount of time teachers assessed student work; (c) use of instructional strategies, (d) use of assessment resources and strategies, and (e) levels of student understanding. Log completion rates varied greatly from week to week, and from teacher to teacher. Results were summarized across logs at the individual teacher level.

Preliminary factor analyses were conducted to better understand how the teacher logs could function as an indicator of fidelity of implementation. The analysis revealed two primary factors (see Table 6). Factor 1, a proxy for general information about implementation (including the frequency of and amount of time teaching the curriculum unit, and the evaluation of and feedback on student work) accounted for 56% of the total variance in the model. Factor 2 identified a useful single-item measure of the minutes per day spent teaching the unit, which is only moderately correlated with days per week teaching it and time spent looking at student work after teaching. In other words, Factor 2 seems to get at the degree of intensity with which class time is focused on the science unit. This factor accounts for 12% of the total variance among the log items. Overall, the alpha for the general implementation factor was 0.81.

Table 6

Teacher Log: Factors Component Matrix

	Component	
	Factor 1	Factor 2
Number of times the science unit was taught/week	.623	.444
Minutes/day > 40 on science unit	.367	.738
Minutes/day (at least 10) on analysis of student work	.678	.307
Provided written feedback on individual student work (notebooks or other) to most students	.833	.100
Used a scoring guide to analyze student work	.783	-.281
Figured out a next instructional step based on student assessment data	.806	-.068
Recorded observations of students during class	.880	-.097
Checked student understandings at the end of an Investigation	.780	-.192
Conducted student self-assessment sessions	.853	-.386

Note. Extraction method: principal component analysis.

Results

Descriptive statistics for each of the study instruments is summarized below. Analysis of the relationships among the measures then follows.

Descriptive Results

Teacher content knowledge. The low reliability of the multiple-choice and completion items precluded any sub score analysis. Results show that teacher scores ranged from a low of 41% (12/29) correct to 86% correct (25/29), with substantial variation in the mean scores (*SD* = 3). Mean, median, and mode scores all centered on 20/29 correct—approximately 70% accuracy. Table 7 shows item difficulty by topic for the final item set.

Table 7

Mean Difficulty for Multiple Choice and Completion items (*n* = 34)

Demonstrated content knowledge topic	Min	Max	Mean	*SD*
Magnetism	.29	1.00	.8088	.15993
Electricity - modified scale	.23	1.00	.6327	.17750
Electromagnetism	.00	1.00	.3676	.35481

Teacher content-pedagogical knowledge: Analysis and use. Although responses to both the analysis and interpretation items and those for next steps of instruction were combined, for the purposes of subsequent analysis, we provide separate descriptives here. Table 8 shows that teacher scores varied widely in both areas.

Table 8

Teacher Scores on Content-Pedagogical Knowledge Tasks (*n* = 34)

Scores	Analysis and interpretation	Next steps
Mean (*SD*)	8.4 (2.4)	6.8 (3.3)
Median	9	7.5
Mode	9	8
Range	6–12	2–13
Total possible	18	18

For *Analysis and Interpretation* items, nearly a third of the teachers left more than half of the items blank, and only a small minority of teachers (5) scored 2 or 3 on at least five of the six items. The remainder of teachers were distributed across the remaining score points. Teachers' scores were highest for the items that focused on magnetism and lowest on electromagnetism items.

Scores from the *Next Steps* items similarly varied greatly. At the low end of the spectrum, four teachers scored a total of only 2 points out of a possible 21 points; at the other end of the continuum, the three highest scoring teachers scored 13 out of a possible 21 points. The average score was 6.7 (*SD* = 3.2), with a range of 2 to 13 points. The scores of the great majority, 80% of the teachers, were clustered around score points "1" and "2" suggesting that most teachers tended to rely on general approaches to subsequent instructional planning. This finding is consistent with other recent studies on teacher pedagogical content knowledge (e.g., Heller, et al., 2004; Heritage & Vendlinski, 2006). These studies also found teachers more proficient at analysis and interpretation than at formulating next steps for instruction and that teachers provided only general information (e.g., review, reteach, or do more investigations/problems) about what they would do next instructionally to support student learning.

Teacher self-report of content knowledge. Table 9 shows teacher perceptions of their own content pedagogical knowledge, which provides a more positive picture than the direct measures. Sampled teachers, on average, reported themselves moderately qualified to teach target concepts in magnetism and electricity. Like the direct measures, teachers' scores were quite varied and showed more knowledge of magnetism than of electromagnetism.

Table 9

Teacher Self-Reported Content Knowledge (*n* = 39)*

How well qualified do you feel to teach 4th grade students about the following topics?			
Topic	Range	Mean	*SD*
Magnetism and electricity	2–5	3.6	0.9
Magnetic forces	2–5	3.6	0.9
How electrical circuits are designed	1–5	3.3	1.1
How electricity can produce magnetic effects	1–5	3.1	1.1
How well do you understand the M&E curriculum unit goals?			
OVERALL		**3.5**	**.8**

Note. *5-pt scale: 1 (*not at all qualified*) to 5 (*highly qualified*), M&E = magnetism and electricity.

Teacher self-report of assessment practices. Table 10 shows how frequently teachers report being engaged with various aspects of the assessment process in teaching the subject science unit. Although there is variation in teachers' responses, in general teachers report usually on a daily basis establishing and communicating their learning goals for students, and regularly, at least weekly, both selecting or developing assessments to address those goals and analyzing/interpreting student work. They more occasionally use specific strategies accompanying the unit to assess and respond to student understandings.

Table 10

Teacher Self Report of Assessment Practices (*n* = 28)

Assessment practice scales	Min	Max	Mean	*SD*
Establish and communicate goals	2.67	5.00	4.4598	.65716
Align assessment with goals	3.00	5.00	4.1143	.60719
Analyze and interpret	2.75	5.00	4.0536	.59844
Use assessment strategies	1.83	4.67	3.4048	.67499

Teacher logs on assessment practices. Table 11 shows descriptive results on how frequently per week teachers taught the curriculum and engaged with various aspects of the assessment process, using aspects that cohered based on the factor analysis of scores (see methodology section above). The data suggest that sample teachers typically taught science three times a week and more than half devoted more than 40 minutes per day to the subject. A similar percentage of teachers reported spending at least 10 minutes on each day they taught science to analyze student work—this was a minimum amount of time suggested by the curriculum developers.

Consistent with the self-report survey data, the log data also suggest that teachers regularly engage with assessment: teachers report providing individual, written feedback, using scoring guides, recoding observations, checking student understandings at the end of investigations and using the data to guide subsequent weekly instruction. Less often teachers provided their students opportunities to engage in self-assessment.

Table 11

Teacher log data on curriculum and assessment use (*n* = 40 teachers)

Time on curriculum and assessment	Mean	*SD*
Number of times science curriculum unit taught/week	2.79	.86
Percentage of logs where teachers reported spending more than 40 minutes/day teaching the science unit	0.57	0.46
Percentage of logs where teachers reported spending at least 10 minutes/day looking at student work	0.60	0.3
Use of Assessments*		
Provided written feedback on individual student work (notebooks or other) to most students	0.98	0.74
Use a scoring or coding guide to analyze student work	0.9	0.84
Figured out a next instructional step based on student assessment data	1.03	0.84
Recorded observations of students during class	1.03	0.93
Checked student understandings at the end of an Investigation	1.17	0.77
Conducted student self-assessment sessions	0.57	0.68

Note. * Number of times per week.

Relationships among constructs

Table 12 in the Appendix displays the correlations among all measures included in this study. While sample size and the reliability of each measure provide strong caveats for any interpretation, study findings with regard to measures of teacher knowledge reveal:

- No relationship between teachers self-report of their content-pedagogical knowledge (i.e., the extent to which they feel qualified to teach specific concepts of units on magnetism and electricity, and direct demonstrations of such knowledge).
- No relationship between basic knowledge of specific concepts of magnetism and electricity and teachers' ability to analyze and suggest next instructional steps based on student responses.

Modest, marginally statistically significant, relationships emerged among the various aspects of assessment practices included in the study, but these were not consistent:

- Teachers who reported more frequently establishing and communicating their learning goals also more frequently reported coordinating their assessments with those goals,
- Teachers who more frequently reported aligning their goals and assessment also tended to report that they more frequently analyzed student and group work and

that they more frequently used a variety of strategies to assess student understanding.

- However, there was no apparent relationship between reported of frequency of establishing goals and frequency of analyzing student progress toward those goals.

Study results show little relationship between teachers' content-pedagogical knowledge and their assessment practices:

- Teachers who were more confident of their content knowledge tended to report more frequent alignment of their instructional goals with their assessments and to report more frequent analysis and interpretation of individual and group work.
- Teachers who scored higher in the analysis and interpretation of student work tended to report more frequent engagement in such analysis.
- No relationship was found between other indicators of assessment quality and practice.

Discussion and Conclusion

We started this report by introducing a model that suggests that quality in formative assessment practice involves the quality of the assessment (i.e., the validity of the inferences) and the quality of the process for using assessment to understand and improve student learning. The study used multiple measures to examine these aspects of quality in practice using available data from a larger study of the effects of adding curriculum-embedded assessments to a hands-on science curriculum. The limitations of study data are obvious: small, non-representative teacher sample, the psychometric quality, and validity of available measures. Nonetheless, findings raise interesting issues with regard to both the conceptualization and measurement of teachers' assessment practice and the nature of current teacher practice.

A first issue relates to the difficulty of getting coherent, valid measures of teachers' assessment practice. On the one hand, the multiple measures used in this study provide a variety of vantage points from which to view teachers' assessment practice. However, the relationships among these multiple measures do not suggest a strong underlying construct. Rather, the general lack of relationship between the quality of teachers' assessment, based on their ability to analyze student understanding and respond with instructional next steps, and teachers' use of the assessment process, based on self report survey data, provides support for our model. Results suggest that it may be important to differentiate teachers' engagement in the process of assessment from the validity of the inferences they are able to draw and use from that process.

However, the inconsistencies between and within different aspects of teachers‘ assessment practices means that it is difficult to be confident about the meaning of our measures. Certainly, we recognize shortcomings in the reliability and validity of the individual measures used here. Available measures confound measurement construct and format—that is, the study relied on teacher self reports for measuring assessment process and used primarily direct measures for examining assessment quality. In light of research showing a disjunction between teachers’ reports of their practice and nuanced observations of it (Ball & Cohen, 1999; Cohen, 1990; Mayer, 1999), it is possible that differences in measurement format as much as substance underlies our findings.

It may also be true that the inconsistencies across the measures used in the study mark the reality of teachers who themselves are somewhere within the process of developing their assessment capacity and thus they engage some aspects of the process more than others, some better than others. Regardless of the reasons for the inconsistencies found in this study, the need for valid measures of practice is worth underscoring. We believe the types of measures used here are typical of those used in assessment research—yet, without sound measures any research findings are suspect.

If we take at face value the possibility that study teachers lie somewhere on a trajectory between novice and accomplished practice, then study results suggest that the process starts with appreciating the value of and trying to engage in assessment. Self-report findings here suggest that teachers are well on their way on these dimensions of the process. Teachers report that they regularly use each part of a systematic assessment process; they establish goals, administer measures, and gather other evidence of student learning and use results for planning and next steps. One interpretation of study findings is that teachers “talk the talk”, but findings on the teacher knowledge and the quality of their assessments suggest that they need help to more fully “walk the walk.”

Keeping in mind the very small sample, the findings on teachers’ content pedagogical knowledge and the accuracy with which sample teachers are able to analyze student responses and suggest next steps are stark. Even though all teachers were experienced and had previously taught the target content, few were able to use student explanations on open-ended items to interpret student understandings and/or misconceptions. Study teachers’ ability to formulate specific next steps for teaching and learning was even more limited. Certainly such findings raise important questions about whether teacher capacity to use assessment to promote learning or to bring the vision of formative assessment to fruition. For example, treatment teachers in the larger study were asked to adopt the following new practices as part of implementing the new curriculum embedded assessment system:

1. Administer assessment tasks on a daily/consistent basis.
2. Use a coding/scoring guide to analyze and interpret the evidence from student data.
3. Provide substantive feedback to students on their performance (i.e., not based on grades but rather conceptual understanding).
4. Administer formal checks at the end of each hands-on investigation to assess specific learning goals.
5. Implement targeted, specific, and appropriate "next-step strategies."
6. Meet in study groups to discuss student work.
7. Analyze and interpret student work, as well as patterns and trends in the data.

This is a daunting list for teachers to incorporate into their practice, and it is important to remember that the data reported here are from the first year of implementation. Given these kinds of challenges, it seems clear that if current federal assessment policies are to nourish student learning, teacher capacity needs serious attention.

We look forward to the next stage of the study, (a) to a larger sample that may confirm or disconfirm these initial findings, (b) data on student learning, through which we will investigate the relationship between teacher practice and the improvement of student learning.

References

American Educational Research Association (AERA), American Psychological Association (APA), and the National Council of Measurement in Education (NCME). (1999). *Test Standards for Educational and Psychological Measurement*. Washington, DC: AERA.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In Darling-Hammond, L, & Sykes, G. (Eds.), *Teaching as the learning profession: Handbook of policy and practice (*pp. 3–32). San Francisco, CA: Jossey-Bass.

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

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*(1), 7–73.

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis*, *12*(*3),* 311–329.

Council of Chief State School Officers (CCSSO). Website on Formative Assessment. Available on: www.ccsso.org/projects/SCASS/Projects/Formative%SFassessment%

Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P., R., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education*, 21(4), 360–389. doi: 10.1080/08957340802347852

Heller, J. I., Daehler, K. R., Shinohara, M., & Kaskowitz, S. R. (2004, April 2). *Fostering pedagogical content knowledge about electric circuits through case-based professional development.* Paper presented at the annual meeting of the National Association for Research in Science Teaching (NARST), Vancouver, Canada.

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom.* Thousand Oaks, CA: Corwin Press.

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. L. (2009). From evidence to action: A seamless process in formative assessment?, *Educational Measurement: Issues and Practice, 28*(3), 24–31.

Heritage, M., & Vendlinski, T. (2006). *Measuring teachers' mathematical knowledge* (CSE Tech. Rep. No. 696). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Herman, J. (2010). Impact of assessments on classroom practice. In E. Baker, B. McGaw, & P. Peterson (Eds.). *The International Encyclopedia of Education,* (3rd ed.). Oxford UK: Elsevier

Herman, J. L., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2006). *The nature and impact of teachers' formative assessment practices* (CSE Tech Rep. No. 703). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hill, H., Schilling, S. & Ball, D. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, *105*(1), 11–30.

James, M., McCormick, R., Black, P., Carmichael, P., Drummond, M., Fox, A., … Wiliam, D. (2007). *Improving Learning how to learn.* United Kingdom: Routledge.

Mayer, D. P. (1999). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis*, *21*(1), 29–45.

National Research Council. (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.). Washington, DC: National Academy Press.

Olson, J. F., Martin, M. O., & Mullis, I. V. S. (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Office of Economic Co-operation and Development (OECD, 2005). *Formative assessment: Improving learning in secondary classrooms*. Paris: OECD publishing.

Shepard, L. (October, 2005). *Formative assessment: Caveat emptor.* Paper presented at the ETS Invitational Conference, The Future of Assessment: Shaping Teaching and Learning, New York, NY.

The Regents of the University of California, Berkeley, Lawrence Hall of Science. (2005). *Full Option Science System (FOSS), magnetism and electricity module, teacher guide.* Nashua, NH: Delta Education.