CRESST REPORT 809

RELATIONSHIPS BETWEEN TEACHER KNOWLEDGE, ASSESSMENT PRACTICE, AND LEARNING-CHICKEN, EGG, OR OMELET

NOVEMBER, 2011

Joan Herman
Ellen Osmundson
Yunyun Dai
Cathy Ringstaff
Mike Timms

CRESST
National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Relationships between Teacher Knowledge, Assessment Practice, and Learning-Chicken, Egg, or Omelet?

CRESST Report 809

Joan Herman, Ellen Osmundson, Yunyun Dai
CRESST/University of California, Los Angeles

Cathy Ringstaff, Mike Timms
WestEd

November, 2011

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

Copyright © 2011 The Regents of the University of California

The work reported herein was supported by prime sponsor number R305B070354 from the US Department of Education to WestEd, grant #5387 s07-091.

The findings and opinions expressed in this report are those of the author(s) and do not necessarily reflect the positions or policies of WestEd, ASK, FOSS, or the US Department of Education.

To cite from this report, please use the following as your APA reference: Herman, J., Osmundson, E., Dai, Y., Ringstaff, C., & Timms, M. (2011). *Relationships between teacher knowledge, assessment practice, and learning-Chicken, egg, or omelet?* (CRESST Report 809). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

Abstract 1
Introduction 1
Methodology 3
  Sample 3
  Study Variables and Instrumentation 5
  Analysis 9
Results 10
  Research Question 1: Teacher Knowledge 10
  Research Question 2: Assessment Implementation 13
  Research Question 3: Student Outcomes 15
  Path Analysis Results 15
Discussion 17
  Research Question 1: What is the quality of teachers' content-pedagogical knowledge? 17
  Research Question 2: What is the relationship between teacher knowledge and assessment practice? 18
  Research Question 3: What is the relationship between teacher knowledge, assessment practice and student learning? 18
Conclusions 19
References 21
Appendix 23

RELATIONSHIPS BETWEEN TEACHER KNOWLEDGE, ASSESSMENT PRACTICE, AND LEARNING - CHICKEN, EGG, OR OMELET?

Joan Herman, Ellen Osmundson, and Yunyun Dai
CRESST/UCLA

Cathy Ringstaff, Mike Timms
WestEd

Abstract

Drawing from a large efficacy study in upper elementary science, this report had three purposes: First to examine the quality of teachers' content-pedagogical knowledge in upper elementary science; second, to analyze the relationship between teacher knowledge and their assessment practice; and third, to study the relationship between teacher knowledge, assessment practice, and student learning. Based on data from 39 teachers, we found that students whose teachers frequently analyzed and provided feedback on student work had higher achievement than students whose teachers spent less time on such activities. Our findings support other research indicating the power of well-implemented formative assessment to improve learning.

Introduction

Spurred by Black and Wiliam's (1998) meta-analysis documenting formative assessment as a powerful classroom intervention, particularly for low achieving students, and supported by researchers and practitioner communities from diverse theoretical perspectives (see reviews by Shepard, 2005; Herman, 2010; James et al., 2007), policymakers across the world are considering formative assessment as a primary approach to educational reform (OECD, 2005; CCSSO, 2008). In the US, billions of dollars have been invested in Race to the Top initiatives that put Common Core State Standards, assessment, and use of data front and center, including over $350 million awarded to two state consortia to develop new standards-based assessment systems. While system development focuses primarily on testing for accountability purposes, the federal assessment grants, for the first time, recognize the importance of formative assessment and of building teachers' capacity to use it.

These are promising developments for pushing formative assessment to fruition in classroom practice. Yet at the same time, recent studies reveal challenges in implementing quality formative practice (Heritage, Kim, Vendlinski & Herman, 2009; Heritage, Jones & White, 2010; Herman, Osmundson, & Silver, 2010*)*; show non-robust results with regard to effects on student learning (Furtak, et al., 2008; Herman et al., 2006*;* Wiley & Ciafolo,

2010); and raise questions about the research base underlying formative assessment (Bennett, 2009). Just as the concept of formative assessment underscores the central role of evidence in effective teaching and learning, so too do policymakers and practitioners need evidence on which to build effective formative practices.

Fundamentally, formative assessment involves the use of assessment to "form" subsequent instruction (Black & Wiliam, 2004) – or as the Council of Chief State School Officers (CCSSO) defines it, "a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes" (FAST/SCASS, 2008). Formative assessment involves knowing what the learning goals are, eliciting evidence of student status relative to the goals, and taking action to close any gap between students' current status and the desired goal(s) (Black & Wiliam, 1998, 2004, 2009; Black, Harrison, Marshall & Wiliam, 2003; Hattie & Timperley, 2007; Heritage, 2010; Sadler, 1989; Shepard, 2005).

Because formative assessment is a dynamic process of evidence elicitation, analysis, and action, it clearly makes demands on teachers' content and pedagogical knowledge. Without such foundational knowledge, teachers' formative assessment may yield faulty decisions that could divert rather than promote student progress. At the same time, there also could be a reciprocal relationship between teachers' use of assessment and their content and pedagogical knowledge. Teachers who engage in formative assessment are continually attuned to and responding to student learning progress. Educators who analyze student learning, consider potential obstacles or misconceptions limiting this learning, and reflect on the effectiveness of prior and subsequent next steps—may well deepen their content and pedagogical knowledge, particularly if such activities occur in the context of professional learning communities (Little, 2003; Stoll et al. 2006)

While the challenge of teachers' content-pedagogical knowledge has been documented (Heritage et al., 2009; Heritage, Jones & White, 2010; Herman et al., 2010), few studies have examined the relationship between such knowledge and teachers' assessment practices, nor examined how teachers' knowledge may moderate the relationship between assessment practices and student learning. The study reported here draws from a larger intervention study in upper elementary science to explore these relationships. Study research questions include:

1. What is the quality of teachers' content-pedagogical knowledge?
2. What is the relationship between teacher knowledge and assessment practice?

3. What is the relationship between teacher knowledge, assessment practice, and student learning?

Methodology

While the larger study is based on two cohorts of teachers across three states who participated in a randomized field study of the effects of incorporating new, curriculum-based assessments into an upper elementary hands-on science curriculum program, this study is based only on the Cohort 1 sample for whom full data are available. Because the validity of the curriculum-based assessments used in the study has been established (Draney et al., 2005), study data provide a good opportunity to examine the role of teacher content-pedagogical knowledge and teacher assessment use in student learning, without any confounding from the quality of the assessment data.

For each cohort, schools in each state (and the teachers within them) were randomly assigned to either treatment (revised program with curriculum-embedded assessments) or control (traditional program) conditions. Treatment teachers participated in two days of summer professional development to orient them to the new curriculum and assessment, follow-up sessions to support the analysis of student work, and a practice year for implementing the curriculum in preparation for the Year Two investigation of treatment impact. Control teachers also participated in a similar amount of summer professional development focused on teaching the original curriculum. All teachers in the study implemented two curriculum units, one on Magnetism and Electricity and the second on either Structures of Life or Water.

Given the focus of the treatment, the study used a variety of methods to collect data on teachers' assessment practices, including teacher surveys, logs, and direct measures of teachers content-pedagogical knowledge. In addition, for the impact study year, measures of student learning were used. Cohort 1 study data are now complete. Cohort 2 teachers are completing their impact study year; thus, full data are not yet available.

Sample

Cohort 1 is comprised of 39 teachers from a southwest state. Table 1—which shows the demographic characteristics of the educators in the study—reveals no major differences between treatment and control teachers.

Table 1

Teacher Demographic Information: Cohort 1

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[a]		
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 ed., science endorsement)	4	5
Grade level taught		
3rd grade	0	0
4th grade	19	20

Descriptor	Control *N*=19	Treatment *N*=20
Years of experience teaching elementary grades		
Average number of years	12.0	8.4
Range of years teaching	1-32	2-25
Years teaching science curriculum unit		
Average number of years	3.0	2.6
Range of years teaching	1-11	2-12
Number of science PD hours in the past 2 years		
Average number of hours	19.6	21.3
Range of hours	4-100	2-80

[a]Teachers may hold multiple credentials; therefore, the total number of credentials represented in the table exceeds the total number of teachers who participated in the study.

Table 1 highlights that the majority of the study's participants were white females who possessed an elementary credential. It is also evident that control teachers were more likely to hold a masters degree than treatment teachers. However, in regards to their average number of years of teaching experience, as well as their experience teaching the science curriculum under study—control and treatment teachers appeared to be more similar.

The project continuation or completion rate for Cohort 1 was high; in fact, of the 39 teachers who began the project in August 2008, 32 teachers (or 82%) remained in the project through its conclusion in June 2010. Most teachers who left the project did so because of changes in teaching assignments to different grades or non-project schools (personal communication, M. Tiu, July, 2010).

Study Variables and Instrumentation

Study variables include teachers' content-pedagogical knowledge, as measured by direct assessment of teachers; teachers' use of assessment, as measured by teacher logs; and multiple, direct measures of student learning.

Teachers' content-pedagogical knowledge (Research Question 1). Measures of teachers' content- pedagogical knowledge were drawn from multiple choice and performance assessments, both of which were administered before the start of the project and at the conclusion of the impact study year. All measures were focused on the magnetism and electricity unit, as this was the only unit implemented by all teachers and the focus served to conserve limited resources, including teacher time. Described more thoroughly in Herman, Osmundson, & Dai, 2010, each of these measures is summarized below.

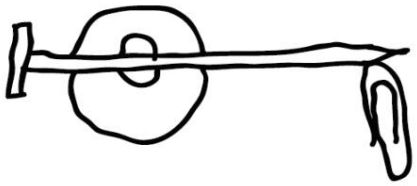
The multiple choice test concentrated on three primary conceptual areas: magnetism, electricity, and electromagnetism; these were the three major concepts addressed in the study curriculum. Although composed of released National Assessment of Educational Progress (NAEP) and state assessment items—test reliability (coefficient alpha) was moderate, at .73 for the total score.

The performance-oriented assessment addressed teachers' capacity to analyze and interpret student responses, a proxy for teachers' pedagogical content knowledge. The structure of these items was as follows:

1. Teachers answered an open-ended content question related to one of the three major concepts;
2. Student responses to the same question were provided, and teachers were asked to analyze and interpret the student responses; and
3. Based on their analysis, teachers were asked to indicate the nature of specific students' understandings and what they would do next to support student progress.

Figure 1 shows a sample item that follows the teacher content survey sequence described above.

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 1. Teacher content survey: Magnetism and electricity module.

Teacher responses to these performance items (Figure 1- parts b & c) were scored based on a 0-3 scale, derived from the defining features of expert ratings of a sample of teacher responses. A score of 0 was used for a non-response or irrelevant response, while 3 reflected a complete and accurate description of student understandings and misconceptions or of next steps for instruction. Three raters participated in the scoring, all experienced science educators who were specially trained on the scoring rubric and familiar with the curriculum module. Pre- and post-test responses were scored together, with scorers blind to testing occasion. Based on a 25% sample of the responses that were double scored, reliability of scoring ranged between 76% agreement to 96% agreement (see Tables 2 & 3).

Table 2

Cohort 1: Pre-survey Inter-rater Reliability, Open-ended Responses

Comparison rater	Rater 1	Rater 2
2	0.96	
	<.0001	
3	0.90	0.86
	<.0001	<.0001

Note. Pearson Correlation Coefficients, $N = 63$.
Prob > |r| under H0: Rho=0.

Table 3

Cohort 1: Post-survey, Inter-rater Reliability, Open-ended Responses

Comparison rater	Rater 1	Rater 2
2	0.86	
	<.0001	
3	0.91	0.76
	<.0001	<.0001

Note. Pearson Correlation Coefficients, $N = 126$.
Prob > |r| under H0: Rho=0.

Table 4 displays score reliabilities for the pre- and post-performance teacher assessment. Results show reasonable reliability for the analysis and interpretation and next step subscales, particularly given the small number of items constituting each. Scores for the content knowledge questions were less reliable than the other two areas, which in part may be due to the small number of items and potential ceiling effects (a total of seven items).

Table 4

Cohort 1 Score Reliabilities for Performance Items on Content Survey

Items	Pre	Post
Content knowledge	0.51	0.48
Analysis and interpretation	0.73	0.81
Next instructional steps	0.79	0.84

Teachers' assessment practices (Research Question 2). Data on teachers' assessment practices were derived from weekly online logs that educators completed as part of the study. The logs were originally designed as a fidelity of implementation measure for the larger study; they asked teachers to report on: (a) how much time they spent teaching the curriculum; (b) what instructional strategies they were using; (3) their use of available assessment tools and strategies; and (4) their evaluation of their students' level of understanding relative to specific learning goals. While log completion rates varied greatly from teacher to teacher—ranging from some teachers completing as few as 2 logs and others completing more than 20—data were available for almost all teachers in the study.

As Table 5 reveals, factor analysis of the log data showed a clear intensity of assessment factor. Moreover, scores on this factor were significantly and positively related to classroom observation measures of assessment quality for the Magnetism and Electricity module. Correlation coefficients revealed a moderately strong relationship -- .75 (for more detail, see Osmundson et al., 2010). Thus, responses to items on this factor were used to construct a measure of teachers' assessment practices for the current study (alpha=.95).

Table 5

Teacher Logs[a] : Principle Component Analysis[b]

Assessment component	Aggregated items (Q_ave=Q_ave/Q1B_SUM)	Factor 1 "assessment"
Used *At a Glance*	Q2A_ave	0.68
Planned & used assessment	Q3A_ave	0.69
Analyzed student work in notebook	Q3b_ave	0.79
Analyzed student work on response sheets	Q3c_ave	0.72
Analyzed observations of students	Q3d_ave	0.74
Recorded and used assessment information on informal data chart	Q3F_ave	0.65

Assessment component	Aggregated items (Q_ave=Q_ave/Q1B_SUM)	Factor 1 "assessment"
Provided feedback to individual student based on analysis of student work	Q3g_ave	0.67
Provided feedback to the entire class based on analysis of student work	Q3I_ave	0.64
Retaught content	Q3K_ave	0.61
Sum (#time curriculum taught/week)	Q1B_SUM	0.76
Average (minutes/day analyzing student work)	Q1D_AVE	0.68

[a]Treatment teachers only.
[b]Rotated Factor Pattern includes data from all three modules.

Raw scores for items loading on the assessment factors were converted into z scores and z scores were used to compute a total "assessment factor" score for subsequent analysis in the path models.

Student outcome measures (Research Question 3). The full study includes pre/post measures of student learning for each of two modules completed for the study. The first is a specially-developed end-of- year (EOY) assessment that addresses core topics within the modules. The second is an end-of-year state assessment in English-language arts, mathematics, and science. However, scoring of the pre/post assessments is not yet complete.

State assessment scores for English language arts (ELA) and math were available for both the year prior and the end of the study year, but the state science assessment was only implemented in grade 4 and thus available only for students for the end of the year in which they participated in the study. Prior year ELA and math scores were used as a covariate, as were available data on student demographics.

The EOY assessment was specially developed by WestEd to address the content of the three modules that were part of the study: magnetism and electricity, water, and structure of life. (Recall that each teacher implemented two modules, magnetism and electricity plus one of the other two). Administered at the end of the study year, the assessment was comprised of 30 multiple-choice questions, 10 on each of the three content areas. Based on the pilot version of the test, reliability was estimated at .76 (KR-20) and standard error of measurement based on KR-20 estimated at 2.57.

Analysis

Descriptive statistics and path analyses were used to examine the study's primary research questions. Because the underlying study involved an assessment intervention,

observed differences in effects on treatment and control teachers also were of interest. That is, because the intervention focused on the availability and use of curriculum-embedded assessment, any treatment effects on teachers might also suggest the impact of assessment use.

Results

Research Question 1: Teacher Knowledge

Multiple-choice pre/post content survey results. The results presented in Tables 6 and 7 display Cohort 1 control and treatment teachers' performance on the multiple-choice post-test, and compares scores before and after participating in the study (i.e., after teaching the ASK/FOSS Magnetism and Electricity unit for two consecutive years). Results show that control and treatment teachers started the study with moderate scores and made gains in all areas of the content survey after two years of study participation. For the control group, the difference in pre/post scores is statistically significant at the .05 level for three of the four scales (magnetism and two electricity concepts). For the treatment group, the difference in the pre/post scores is statistically significant for *all* four scales, at the .05 level. For both groups, scores started the lowest and increased most on items relating to electromagnetism.

Table 6

Cohort 1 Pre/Post Magnetism and Electricity Content Survey Scores: Control Teachers

	Pre/post multiple choice scores control teachers						
Investigation	*N*	Pre	Post	Pre/post	*df*	*t* value	Pr >|t|
Magnetism	11	0.69	0.83	0.14	10	1.85	0.093
Electricity 1	11	0.68	0.88	0.20	10	4.9	0.001
Electricity 2	11	0.65	0.87	0.22	10	3.36	0.007
Electromagnetism	11	0.50	0.82	0.32	10	1.64	0.1319

Table 7

Cohort 1 Pre/Post Magnetism and Electricity Content Survey Scores: Treatment Teachers

	Pre/Post multiple choice scores treatment teachers						
Investigation	*N*	Pre	Post	Pre/post	*df*	*t* value	Pr >|t|
Magnetism	13	0.75	0.96	0.21	12	5.45	0.000
Electricity 1	13	0.68	0.90	0.22	12	5.42	0.000
Electricity 2	13	0.64	0.90	0.26	12	4.98	0.000
Electromagnetism	13	0.35	0.92	0.58	12	6.04	<.0001

Note. Unique teacher IDs were used to match teachers' pre/post scores. Not all teachers completed pre/post content surveys, due to scheduling conflicts, hence the lower number of teacher scores reported than study participants.

In general, we did not find statistically significant differences in teachers' post-test knowledge on the multiple-choice test as a function of treatment condition (control vs. treatment). The results for the first subscale (magnetism) were an exception; in fact, when controlling for pre-test performance, treatment teachers outperformed control teachers on the multiple choice post-test items on magnetism.

Content-pedagogical performance assessment results. Pre- and post-teacher scores for the content-pedagogical performance assessment are shown in Table 8. For both groups, teachers' initial scores in content were modest, similar to the multiple-choice results, but scores on performance analysis and interpretation of student work and knowledge of instructional next steps were quite low, achieving on average only 29% to 37% of total possible points. While treatment teachers' scores appeared slightly higher at the beginning of the study, the differences are not statistically significant.

Similar to the multiple-choice trends, both groups also improved their scores on the content-pedagogical performance assessment after teaching the Magnetism and Electricity Module for two consecutive years. For both groups, the largest gain was in the area of instructional next steps.

Table 8

Cohort 1: Pre/Post Content-Pedagogical Performance Scores

	Control *N*=13				Treatment *N*=16			
Items	Pre	% Correct	Post	% Correct	Pre	% Correct	Post	% Correct
Content[a]	4.1	58.57	5.4	77.14	4.4	62.86	6.6	94.20
Analysis and interpretation[b]	6.5	30.95	10.8	51.43	7.8	37.14	14.7	70.00
Instructional next steps[c]	6.0	28.57	11.4	54.29	7.6	36.19	14.7	70.00

[a]Scale =1 (correct), 0 (incorrect), 7 possible points; [b]Scale range 0 – 3 (see scales above), 21 possible points; [c]Scale range 0 – 3 (see descriptions above), 21 possible points.

Regression analyses using the pre-survey score for each question type as a covariate showed a statistically significant treatment effect for teachers' content-pedagogical knowledge. For all three areas – content, analysis and interpretation, and instructional next steps – treatment teachers outperformed control teachers. Table 9 displays results of the regression analyses.

Table 9

Cohort 1 Regression Analysis for Pre/Post Content Survey

Variable	*df*	Parameter estimate	*SE*	*t* value	Pr > |t|
Post content					
Intercept	1	5.15	0.58	8.96	<.0001
Pre content	1	0.06	0.12	0.46	0.65
Treatment	1	1.20	0.39	3.06	0.01
Post analysis & interpretation					
Intercept	1	8.77	1.44	6.11	<.0001
Pre analysis & interpretation	1	0.31	0.18	1.73	0.10
Treatment	1	3.50	1.20	2.92	0.01
Post next step					
Intercept	1	9.16	1.46	6.27	<.0001
Pre next step	1	0.37	0.19	1.98	0.06
Treatment	1	2.73	1.31	2.08	0.05

Research Question 2: Assessment Implementation

Log results. Table 10 summarizes the descriptive results using the teacher as the unit of analysis and mean scores for each item over the course of the unit. While average responses were generally similar between treatment and control teachers, it is noteworthy that treatment teachers spent significantly more time looking at student work and were more likely to report spending more than 10 minutes a day in doing so than control teachers.

On average, the results reported in Table 11 indicate that teachers used the study modules three times a week and, on each of these days, reported spending 5-10 minutes analyzing student' work on the modules. Teachers reported that they most frequently analyzed their observations of students, roughly half the days they taught the modules. Other assessment activities – provision of individual feedback to students, analysis of student written work in notebooks or response sheets – occurred with relatively less frequency. While there is considerable variability in these scores, it is consistent with the variability in the number of times a week teachers reported using the modules – that is, if teachers used the modules for science four times a week, there was the possibility of engaging in each assessment activity four times, while if the modules were used three times a week, the frequency of potential assessment use would be reduced accordingly.

Table 10

2009-2010: Cohort 1 Teacher Log Data Descriptive Results (All Modules)

Teacher log questions	Cohort 1: Control *N*=14 teachers	Cohort 1: Treatment *N*=16 teachers
Number of times/week used modules	2.9 (0.9)	3.1 (0.8)
Assessment time		
Average minutes/day looking at student work	5.9 (5.8)	10.8 (6.9)
Percentage of logs where teachers reported spending more than 10 minutes/day looking at student work	20% (0.3)	43% (0.4)
Use of assessments[a]		
Provided feedback to individual students based on analysis of student work	1.2 (0.8)	1.2 (0.7)
Analyzed observations of students	1.6 (1.1)	1.6 (0.8)
Checked on student understandings at the end of an investigation	1.4 (0.9)	0.8 (0.5)
Engaged students in self-assessment of science learning	1.0 (1.1)	0.7 (0.5)
Analyzed student work in science notebooks	1.1 (1.1)	1.4 (0.8)
Analyzed student work on student response sheets	1.3 (0.8)	1.1 (0.8)

[a]Scale = Number of times/week teacher reported engaging in activities.

Table 11 shows factor scores on assessment implementation for control and treatment teachers. Results suggest that treatment teachers' logs revealed significantly greater implementation of assessment than did control teachers, as judged by time spent analyzing student work and use of various strategies.

Table 11

Assessment Implementation Factor Scores for Cohort 1Treatment and Control Teachers

Group	Assessment Implementation Scores	*N*
Control	-0.82	14
Treatment	-0.07	16
Difference treatment/control	0.75*	

*Statistically significant at the alpha <0.05 level.

Research Question 3: Student Outcomes

Entering abilities. Student state assessments in reading results at the end of the year, which were taken prior to the study, provide a gauge of students' entering ability. Table 12 summarizes students' mean scale scores; these scores indicate that, on average, students participating in the study were performing at the level of "meeting state standards" in reading.

Table 12

Third Grade State Assessment Results in Reading for Cohort 1 Treatment and Control Students

Group	*N*	Mean	*SD*
Treatment	400	470.10	51.32
Control	314	465.50	51.95

State assessment and end-of-year science results. Table 13 summarizes classroom results for the end-of-year science test (developed especially for the study) as well as the state assessment in science. Both tests were administered to students at the completion of the study year. The data show that, on average, students achieved about 60% correct on the end-of-year measure.

Table 13

End-of-Year Science Assessment Scores for Cohort 1 Treatment and Control Classrooms

	Treatment students			Control students		
Test	*N*	Mean	*SD*	*N*	Mean	*SD*
End-of-year science assessment	410	19.18	4.45	323	18.28	5.04
State assessment: Science	435	533.00	54.47	344	522.40	51.82

Path Analysis Results

Path analysis using the combined data from the treatment and control groups was utilized to test the relationships among and between teachers' initial content-pedagogical knowledge, assessment practices, and student learning. The path model controlled for students' entering ability and predicted that teachers' assessment practices and student learning would be directly related, but that teachers' content knowledge would have only an

indirect relationship to student learning through its influence on teachers' assessment practices. That is, teachers with higher content-pedagogical knowledge were expected to engage in more use of assessment, and greater use of assessment was expected to be positively related to student learning, but no direct relationship was expected between teachers' content-pedagogical knowledge and student learning.

Because the teacher measures were highly related and produced similar models, a composite measure of teachers' content-pedagogical knowledge was used, which combined results across the multiple choice and performance assessment measures at the start of the study (i.e., pre-test). Models were tested using both the end-of-year student assessment and state science assessment results as the indicator of student learning. Moreover, in order to examine the extent to which teachers' use of assessment might positively influence their content-pedagogical knowledge, analyses also examined the relationship between changes in teachers' content-pedagogical knowledge and assessment practices. All models were tested at the teacher level and used students' prior year (3^{rd} grade) state reading scores to control for any differences in entering student ability.

Figure 2 displays the standardized path coefficients evident in the relation between the students' entering ability, based on standardized test results in reading; a composite measure of teachers' content-pedagogical knowledge at the start of the study; teachers' assessment use, as measured by weekly logs; and student performance on the end-of-year science assessment. The path coefficients show how much a change of one standard deviation in a prior variable would produce in standard deviation units of the subsequent variable. Results show that, after controlling for students' prior ability, teachers' assessment use is significantly and positively related to students' end-of-year performance. A change of one standard deviation in teachers' assessment use scores is associated with a change of .26 standard deviation in students' performance. Teachers' content-pedagogical knowledge has no direct relationship to student learning, but indirectly affects it through a marginally significant relationship to teachers' assessment use. That is, stronger content-pedagogical knowledge is marginally and positively related to teachers' assessment use, which is positively related to student learning. Bentler's Comparative Fit Index show a good fit for the model displayed (CFI=.9865). Analyses are ongoing and use students' end-of-year scores on the state science assessment.

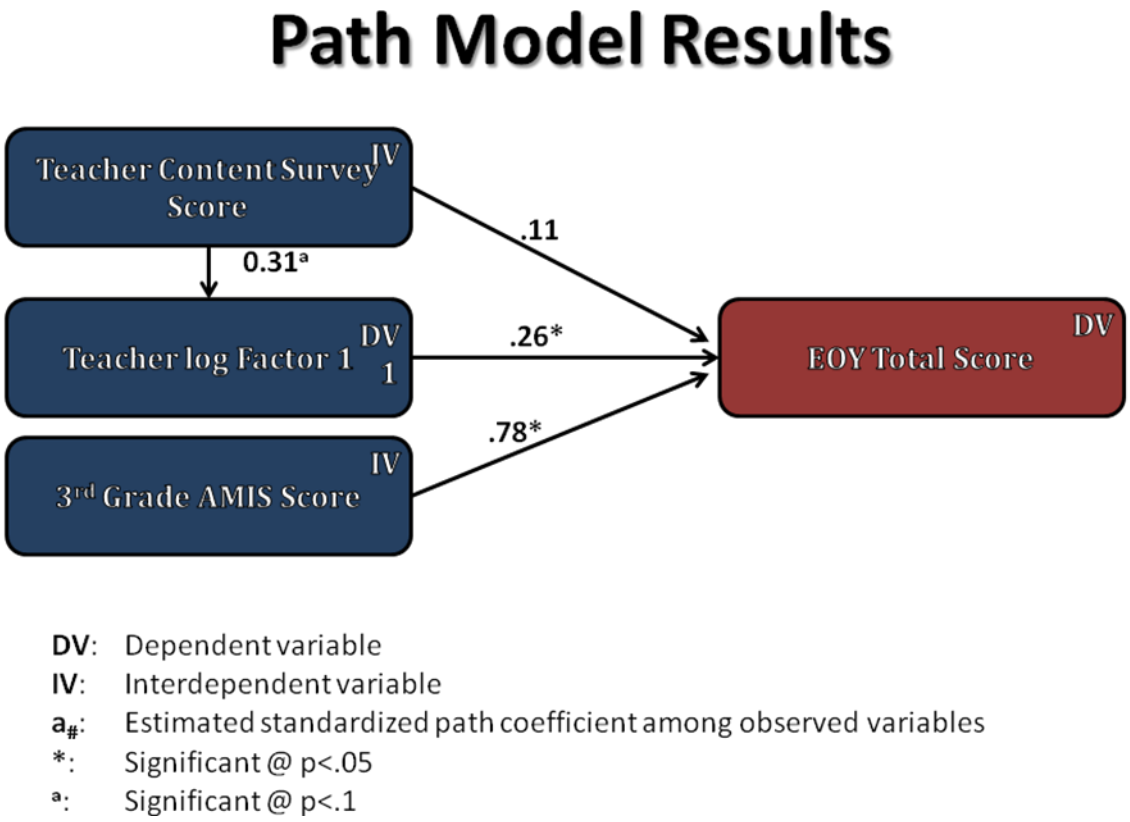


Figure 2. Standardized path coefficient model.

Study analyses found no relationship between changes in assessment use and changes in teacher practices. That is, teachers who spent more time in assessing and responding to students' work did not gain more content-pedagogical knowledge over the course of the study than those who spent less time.

Discussion

This paper started with three research questions. We end it by summarizing our findings with regard to each question and then consider implications.

Research Question 1: What is the quality of teachers' content-pedagogical knowledge?

The study used multiple-choice and performance assessments to measure teachers' content and pedagogical knowledge. The multiple choice test was drawn from publically available items on magnetism, electricity, and electromagnetism intended for elementary students (which sets an admittedly low bar for teachers' content knowledge). At the beginning of the study teachers scored from 35% to 75% correct. By the end of the study,

after having taught the study curriculum twice, performance ranged from 82% to 92% correct.

While results of the performance assessment showed the same positive trends from pre- to post-study, the levels of performance were less promising, particularly for tasks on the analysis and interpretation of student work and knowledge of instructional next steps, where scores were quite low. Scores on the pre-test ranged from 29% to 37% of total possible points and from 45% to 54% on the post-test. For both treatment and control groups, the largest gain was in the area of instructional next steps.

The fact that the performance assessment actually engaged teachers in formative assessment (i.e., analyzing student work and identifying implications for subsequent instruction) also bears directly on participating teachers' assessment capacity. Consistent with prior literature (Herman et al., 2010; Heritage et al. 2009), results suggest limited teacher capacity.

Research Question 2: What is the relationship between teacher knowledge and assessment practice?

Because the study treatment focused on systematically embedding formative and end-of-investigation assessments in a hands-on science curriculum and encouraging teachers to regularly analyze student work, observed differences in the relative gains in content-pedagogical knowledge for treatment relative to control teachers were suggestive of the effects of sound assessment use on teacher knowledge. That is, regression analyses controlling for teachers' pre-study performance showed a statistically significant treatment effect for teachers' content-pedagogical knowledge at the completion of the study, which might be related to the treatment's stronger focus on assessment and the quality of those assessments. For all three areas – content, analysis and interpretation, and instructional next steps – treatment teachers outperformed control teachers.

However, path analyses examining the relationship between teachers' assessment use and changes in teachers' content knowledge revealed no statistically significant relationships. Higher use of assessment, as measured by teachers' responses to weekly logs, was not associated with stronger teachers' content-pedagogical knowledge.

Research Question 3: What is the relationship between teacher knowledge, assessment practice and student learning?

Path analysis results supported Black and Wiliam's conclusions (1998) and the paper's hypothesis about the relationship between teachers' use of assessment in instruction and

student learning. Controlling for students' entering ability, as measured by standardized test scores in reading, the manner in which teachers utilized assessment was positively related to student learning outcomes. More use of assessment was associated with higher student performance at the end of the study. As expected, teachers' initial content-pedagogical knowledge, as gauged by multiple measures administered at the start of the study, showed no direct influence on student learning, but was marginally related to teachers' assessment use. Admittedly, however, the relationship was of little practical significance. The overall model fit was very high for these analyses.

Conclusions

Study findings reinforce the power of formative assessment, or at least one important element of it: Students whose teachers spend more time and who more frequently engage in analyzing and providing feedback on student work achieve higher learning than students whose teachers spend less time and who less frequently do so. Teachers' attention to student learning as evidenced in classroom work—whether through observations of students in classroom discussions or analyses of student responses in science notebooks, other written responses, or end-of-investigation assessments—is associated with higher student performance.

The strength of this relationship is striking in light of the weaknesses in teachers' initial content-pedagogical knowledge, as documented in pre-test scores for this study. It seems obvious that sound formative assessment practice requires adequate content-pedagogical knowledge. In other words, it is hard to imagine how teachers with weak knowledge of subject matter content and of the nature of students' progression through the content can appropriately analyze student work, or make appropriate decisions for next steps. Path analysis results from this study weakly support this supposition, as teachers' content knowledge showed an indirect relationship with student learning through teachers' use of assessment.

That both treatment and comparison teachers showed substantial pre- to post-intervention gains may be at least partially due to a testing effect (the same assessment was given pre-intervention and two years later, at the end of the study). However, since the treatment group showed substantially higher content-pedagogical knowledge at the end of the study, it is possible that the use of sound assessment tools, as embedded in the intervention, may contribute to the development of stronger teacher knowledge. While the path model examining changes in teachers' knowledge did not show this connection, the small sample size, particularly of the treatment group, may have been an obstacle. The addition of Cohort 2

data from our study will broaden the sample for addressing these issues and form clear next steps for us.

We conclude with the same chicken or egg problem that we used in titling the paper. Teachers' use of formative assessment benefits student learning, as the findings reported here substantially document. Yet, effective formative assessment places heavy demands on teachers' content-pedagogical knowledge—knowledge that may be spotty based on the present' study's findings as well as other research. Can analyzing student responses to sound assessments help teachers strengthen their content-pedagogical knowledge? Is a minimal level of knowledge necessary for effective assessment use that benefits learning? What are optimal approaches for developing teachers' capacity in these areas? The current study raises possibilities, but the chicken-egg issue remains unresolved. The omelet remains on the table, which continues to be a popular entree in current policy initiatives. The present study's results underscores both the potential and challenge of bringing these initiatives to fruition.

References

Bennett, R. (2009). *Formative assessment: A critical review.* Presentation at the University of Maryland, College Park, MD.

Black, P. J., & Wiliam, D. (2009). Developing the theory of formative assessment. In *Educational Assessment, Evaluation, and Accountability, 21*(1), 5-31.

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

Black, P. J., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*, 139-48.

Black, P. J., & Wiliam, D. (2004a). Classroom assessment is not (necessarily) formative assessment (and vice-versa). In M. Wilson (Ed.). *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education* (Pt. 2, pp. 183-188). Chicago, IL: University of Chicago Press.

Black, P., & Wiliam, D. (2004b). The formative purpose: Assessment must first promote learning. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability* (pp. 20-50). Chicago, IL: University of Chicago Press.

Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Buckingham, UK: Open University Press.

Council of Chief State School Officers. (2008). *Attributes of effective formative assessment.* Washington, DC: Author.

Draney, K., Galpern, A. & Wilson, M. (2005, November). *Designing and using an embedded assessment system to track student progress.* Presented at the National Science Teachers Association conference, Chicago, IL.

Formative Assessment for Students and Teachers (FAST) and the State Collaborative on Assessment and Student Standards (SCASS). (2008, October). *Attributes of effective formative assessment.* Paper prepared for the Formative Assessment for Teachers and Students State Collaborative on Assessment and Student Standards of the Council of Chief State School Officers, Washington, D.C.

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

Hattie J., Timperley H. (2007). The power of feedback. *Review of Educational Research 77*, 81–112.

Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* Report prepared for the Council of Chief State School Officers.

Heritage, M., Jones, B., & White, E. (2010, April). *Supporting teachers' use of formative assessment evidence to plan the next instructional steps*. Presentation at the annual meeting of the American Educational Research Association, Denver, CO.

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

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).

James, M., Black, P., Carmichael, P., Drummond, M. -J., Fox, A., MacBeath, J., et al. (2007). *Improving learning how to learn in classrooms, schools and networks*. London, UK: Routledge.

Little, J. W. (2003, August). Inside teacher community: Representations of classroom practice. *Teachers College Record, 105*(6), 913-945.

OECD. (2005), *Formative assessment: Improving learning in secondary classrooms*. Paris, France, Author.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-140.

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

Stoll, L., Bolam, R., McMahon, A., Wallace, M., & Thomas, S. (2006). Professional learning communities: A review of the literature. *Journal of Educational Change, 7*(4), 221-258.

Wylie, E. C., & Ciafalo, J. (2010, April). *Documenting, diagnosing, and treating misconceptions: Impact on student learning.* Paper presented at the American Education Research Association, Denver, CO.

Appendix