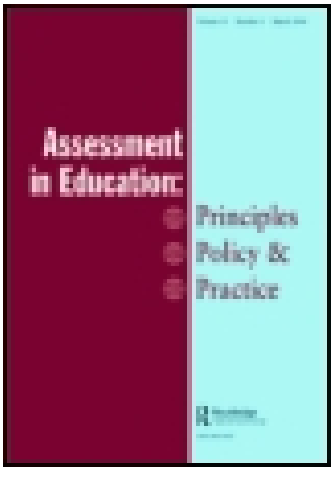

Assessment in Education: Principles, Policy & Practice
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/caie20

Investigating the dynamics of formative assessment: relationships between teacher knowledge, assessment practice and learning

Joan Herman[a], Ellen Osmundson[a], Yunyun Dai[a], Cathy Ringstaff[b] & Michael Timms[c]

[a] Graduate School of Education and Information Studies, CRESST, University of California, Los Angeles, CA, USA

[b] WestEd, Redwood City, CA, USA

[c] Australian Council for Educational Research (ACER), Melbourne, VIC, Australia

Published online: 23 Mar 2015.

To cite this article: Joan Herman, Ellen Osmundson, Yunyun Dai, Cathy Ringstaff & Michael Timms (2015): Investigating the dynamics of formative assessment: relationships between teacher knowledge, assessment practice and learning, Assessment in Education: Principles, Policy & Practice, DOI: 10.1080/0969594X.2015.1006521

To link to this article: http://dx.doi.org/10.1080/0969594X.2015.1006521

Assessment in Education: Principles, Policy & Practice, 2015
http://dx.doi.org/10.1080/0969594X.2015.1006521

Routledge
Taylor & Francis Group

Investigating the dynamics of formative assessment: relationships between teacher knowledge, assessment practice and learning

Joan Herman[a]*, Ellen Osmundson[a], Yunyun Dai[a], Cathy Ringstaff[b] and Michael Timms[c]

[a]*Graduate School of Education and Information Studies, CRESST, University of California, Los Angeles, CA, USA;* [b]*WestEd, Redwood City, CA, USA;* [c]*Australian Council for Educational Research (ACER), Melbourne, VIC, Australia*

(*Received 10 January 2013; accepted 8 January 2015*)

This exploratory study of elementary school science examines questions central to policy, practice and research on formative assessment: What is the quality of teachers' content-pedagogical and assessment knowledge? What is the relationship between teacher knowledge and assessment practice? What is the relationship between teacher knowledge, assessment practice and student learning? Drawing on multiple measures, hierarchical linear modelling and path analysis, results suggest that despite weaknesses in teachers' content-pedagogical and assessment knowledge, teachers' formative assessment practices are positively related to student learning. Relationships between teachers' knowledge and assessment practices are mixed. Findings underscore both the potential and challenge of bringing effective formative practice to fruition as well as the need for continued research.

Keywords: formative assessment; teacher content-pedagogical knowledge; assessment practices

Spurred by Black and Wiliam's (1998a) review 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 Herman, Osmundson, & Silver, 2010; James et al., 2007; Shepard, 2005), policy-makers across the world are considering formative assessment as a primary approach to educational reform (Organisation for Economic Co-operation and Development [OECD], 2005; Council of Chief State School Officers [CCSSO], 2008). In the United States, 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 centre, including $330 million awarded to two state consortia to develop new standards-based assessment systems (United States Department of Education, 2010). While system development focuses primarily on testing for accountability purposes, the federal assessment grants, for the first time, recognise the importance of formative assessment and of building teachers' capacity to use it.

*Corresponding author. Email: herman@cse.ucla.edu

These are promising developments for pushing formative assessment to fruition in classroom practice in the United States. Yet, at the same time, recent studies reveal challenges in implementing quality formative practice (Heritage et al., 2009; Heritage, Jones, & White, 2010; Herman et al., 2010); show non-robust results with regard to effects on student learning (Furtak et al., 2008; Herman, Osmundson, Ayala, Schneider, & Timms, 2006; Wylie & Ciafalo, 2010); and raise questions about the research base underlying formative assessment (Bennett, 2011). In a recent quantitative meta-analysis, Kingston and Nash (2011) document statistically significant effects for formative assessment, but effect size findings are substantially lower than that observed in Black and Wiliam's seminal (1998a) review. Results also showed variation by subject matter; observed effect sizes, for example, were lowest for science (.09), the topic of this study. We hold that just as the concept of formative assessment underscores the central role of evidence in effective teaching and learning, so too do policy-makers 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, 2004a), or as the United States' CCSSO (2008) defines it, 'a process used by teachers and students during instruction that provides feedback to adjust on-going teaching and learning to improve students' achievement of intended instructional outcomes' (FAST/SCASS, 2008). Formative assessment involves being clear on learning goals, eliciting and analysing evidence of student status relative to the goals, providing feedback and taking action to close any gap between students' current status and the desired goal(s) (Black & Wiliam, 1998b, 2004b, 2009; Black, Harrison, Lee, Marshall, & Wiliam, 2003; Hattie & Timperley, 2007; Heritage, 2010; Sadler, 1989; Shepard, 2005).

A dynamic process of evidence elicitation, analysis and action, formative assessment clearly makes demands on teachers' content and pedagogical knowledge. As defined by Shulman and others, such content-pedagogical knowledge[1] involves the practical amalgam of both content and pedagogical knowledge that teachers need to guide effective teaching and learning. In addition to knowledge of the structure, representation and key ideas of specific subject matter domains, content-pedagogical knowledge involves knowledge of how best to develop student understanding in specific content areas, the difficulties and misconceptions that students may face in such development, as well as the specific teaching and assessment strategies teachers need to response to students' learning needs (Ball, Thames, & Phelps, 2008; Heritage et al., 2009; Shulman, 1986, 1987). And in fact, in our view, formative assessment may be seen as teachers' content and pedagogical knowledge in action, the knowledge teachers need in specific classroom contexts to plan instruction, know when and how to elicit student understanding, analyse and interpret student work relative to likely misconceptions and/or obstacles and how to take immediate steps to close the gap between where students are and where they need to be (Heritage, 2010).

Absent or with limited foundational content-pedagogical knowledge, teachers' formative assessment may yield faulty decisions that could delay rather than promote student progress. At the same time, we postulate that there may be a reciprocal relationship between teachers' use of assessment and their content-pedagogical knowledge. That is, to be effective, teachers who engage in formative assessment must be continually attuned to and respond to student learning progress, and teachers' engagement in the analysis of student learning could lead to increases in their

content and pedagogical knowledge. Research, in fact, suggests that teachers who analyse 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, Bolam, McMahon, Wallace, & Thomas, 2006).

While some studies have posited and found a direct relationship between teachers' content-pedagogical knowledge and student learning (see, for example, Hill, Rowan, & Ball, 2005), we view classroom opportunities to learn and the quality of teachers' instructional practices as key intervening variables between teacher knowledge and student learning. That is, teachers' content-pedagogical knowledge is an attribute that influences how and how well teaching, learning and assessment are instantiated in classroom practice, and it is the quality (and content) of these practices that directly affects student learning (see Fennema & Franke, 1992; Wilkins, 2008). We see the influence of teachers' knowledge (content, pedagogical and assessment) in the quality with which teachers enact formative assessment through establishing learning goals, eliciting and interpreting evidence of student learning and providing targeted and specific feedback. Our hypothesis is that each of these teacher actions is influenced and moderated by the teacher's knowledge base.

The challenge of teachers' content-pedagogical knowledge has been documented (Heritage, Kim, Vendlinksi, & Herman, 2009; Heritage et al., 2010; Herman et al., 2010), but 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. In this study, research questions include:

- What is the quality of teachers' content, pedagogical and assessment knowledge?
- What is the relationship between teacher knowledge and assessment practice?
- What is the relationship between teacher knowledge, assessment practice and student learning?

Methodology

The study reported here draws from cohort 1 of a large, randomised field study of the effects of incorporating new, curriculum-based assessments into an upper elementary, hands-on science curriculum programme. Created through a rigorous development and validation process funded by the United States National Science Foundation, the curriculum-based assessments included both embedded tools and strategies for diagnosing and supporting student learning through teacher observation, analysis of student work and student self-reflection and psychometrically sound, end-of-investigation assessments that could be used for formative and/or summative purposes (see Long & Malone, 2006; Wilson & Draney, 2004).

Schools (and the teachers within them) were randomly assigned to either treatment (revised curriculum with embedded assessments) or control (traditional curriculum) 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. Both treatment and control teachers also attended a third day of professional development after the pilot year, to reorient them to study requirements. (We retain the treatment/control distinction here because we regard it as a potential indicator of assessment impact.) All teachers in the study implemented two curriculum units, one on Magnetism and Electricity and the second on either Structures of Life or Water, depending on their school district's mandate.

The study used multiple methods to collect data on teachers' assessment practices, including weekly logs and direct measures of teachers' content-pedagogical knowledge. In addition, study instrumentation included multiple measures of student learning. We summarise study methodology in the pages that follow; further details can be found in Osmundson, Herman, and Dai (2011).

Sample

Teacher sample

The initial study sample comprised 39 fourth grade, volunteer teachers from a southwest state, 20 treatment teachers and 19 control teachers. Analysis of teacher demographic and background information showed no major differences between the two groups of teachers. Study participants were primarily white females holding general elementary credentials. They averaged 8–10 years of teaching experience, 3 years of prior experience teaching the subject curriculum units and 20 hours of participation in professional development on science over the past two years. Over half possessed masters or higher degrees.

Of the 39 teachers who began the project in August 2008, 30 teachers (or 77%) remained in the study through its conclusion in June 2010. Most teachers who left the project did so because of changes in teaching assignments to different grades or because of moving to non-project schools.[2] Student learning data and teacher log implementation data were available for all 30 teachers, but only 24 teachers completed pre- and post-measures of teacher knowledge, as described below. Our small sample sizes led us to use all available data for each study analysis, which depending on the measures involved, ranged from a low of 24 teachers to a high of 30 teachers, as noted in the tabled data below.

We note the limitations that this sample size places on the power and generalisability of the study. Hypothesised relationships have to be very strong to reach conventional levels of statistical significance, and study generalisability is limited by both original sample size and attrition. Missing data, which as noted vary from one data-set to another, further constrains study representativeness and generalisability.

Student sample

Table 1 shows the demographic characteristics of the students included in the study, based on available data for the students of the sampled teachers who completed the study. Data are combined for treatment and control students, as no statistically significant differences were found between the two groups. The table shows a study sample that is balanced by gender and ethnically diverse, with a sizable representation of students of low economic means, as indicated by the 41% of students who

Table 1. Demographic characteristics of student sample. Total initial sample of 877 students, 39 teachers.

Demographic characteristics	Students with available demographic data	n	Percent
Ethnicity			
Hispanic	792	285	36
White	792	417	53
Other (Asian, Black, American Indian, Alaskan Native)	792	86	11
Qualify for free and reduced price lunch	877	356	41
Language status: English language learner	876	130	15
Gender: Male	793	394	50

qualify for the United States' federal free and reduced price lunch programme. Further, 15% of the students are classified as English learners, indicating that they are not yet fully proficient in English.

Study variables and instrumentation

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

Teachers' content-pedagogical knowledge

Teachers' content-pedagogical knowledge, as noted earlier, represents the knowledge that teachers need for effective teaching and learning. It involves understanding of specific content domains as well as knowledge of how student learning is likely to develop in those domains, the challenges and misconceptions students are likely to face along the way, as well as the strategies and learning environments that can best support student progress (Ball et al., 2008; Heritage et al., 2009; Shulman, 1986, 1987). We used multiple measures to capture teachers' content-pedagogical knowledge, including multiple-choice and performance assessments, both of which were administered before the start of the project and at the conclusion of the study year. Our performance measure, as described further below, asked teachers to demonstrate their content understanding and also elicited teachers' content-pedagogical knowledge by asking teachers to analyse student work and suggest next steps to close identified misconceptions and/or gaps in understanding. Due to resource constraints and power issues, all measures were focused on the one unit that all study teachers implemented: magnetism and electricity.

The 30-item multiple-choice content test addressed three topic areas – magnetism, electricity and electromagnetism – the three major concepts addressed in the study curriculum. Composed of released NAEP and state assessment items, test reliability (coefficient alpha) for the total score was moderate, at .73. The expectation was that teachers would have mastery/expert-level knowledge of the material, as the items were fourth grade-level concepts, aligned with the magnetism and electricity module of study.

The performance-oriented assessment addressed teachers' capacity to analyse and interpret student responses, a proxy for teachers' pedagogical-content and

assessment knowledge. The assessment included seven open-ended item sets of three items each, for a total of 21 items. Each item set included a 'content' item, an 'analysis and interpretation' item and a 'next instructional steps' item, structured as follows:

(1) Teachers answered an open-ended content question related to one of the three major concepts (content) addressed by the curriculum (magnetism, electricity and electromagnetism);
(2) Student responses to the same question were provided, and teachers were asked to analyse and interpret the student responses (analysis and interpretation);
(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 (next instructional steps).

Two of these item sets focused specifically on magnetism, three on electricity and two on electromagnetism.

Teacher responses to the first, 'content knowledge' item in each set were scored right or wrong (1, 0), which across the seven sets of items meant a total possible score of seven. Parts 2 and 3 – 'analysis and interpretation' and 'instructional next steps' items – were each scored using a 0–3 scale (see Figure 1, parts b and c), yielding a total possible score of 21 on each of these components. The rubrics were derived from an expert's analysis of the defining features of a sample of high, medium and low responses. A score of zero was used for a non-response or irrelevant response, while a score of three reflected a complete and accurate description of student understandings and misconceptions or of next steps for instruction, a score of two reflected a generally accurate but incomplete description (lacking detail) and a score of one reflected a minimal or broad, vague response.

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. To establish inter-rater reliability, approximately 25% of the teacher performance assessments were randomly selected for double scoring, including six pre-tests and nine post-tests. (Each test included a teacher's response to all 21 constructed response items.) As Table 2 shows, exact agreement for each pair ranged from 80% to 90% and agreement including plus or minus one score point was essentially 100%.

Table 3 displays score reliabilities (based on coefficient alpha) for each component of the pre- and post-teacher performance assessment, without regard to the scorer variation noted above. Results show reasonable reliability for two of the three components, 'analysis and interpretation' and 'next instructional steps', particularly given the small number of items constituting each. Scores on the content component questions were less reliable than the other two scales, which may be in part due to the small number of items and limited range of scores (recall that these items were scored 1–0, for a total of seven possible points).

A total teacher content-pedagogical knowledge score also was created for the path analysis described below. The total score combined all available evidence on teacher content and pedagogical knowledge: raw scores from the three performance assessment components plus scores from the multiple-choice content assessment.

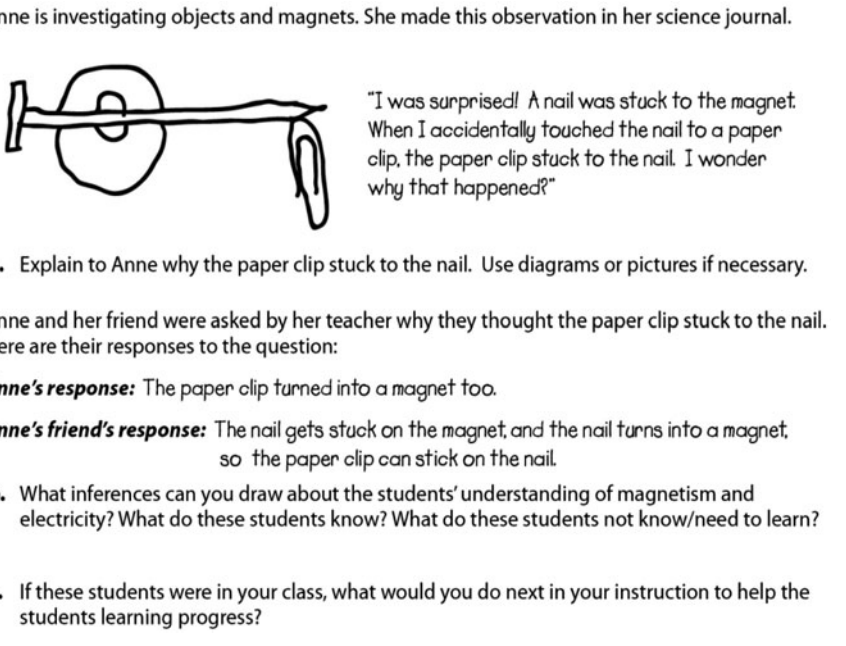

Figure 1. Teacher content survey: Magnetism and electricity module.
Note: Figure 1 shows a sample item that follows the teacher content survey sequence described above.

Table 2. Rater agreement for teacher performance assessment.

	Pre assessment ($n = 6$ teacher test forms)			Post assessment ($n = 9$ teacher test forms)		
Rater pair	*n* (items)	Exact agreement (%)	Off ±1 (%)	*n* (items)	Exact agreement (%)	Off ±1 (%)
Raters 1 & 2	63	95	5	168	84	15
Raters 2 & 3	84	85	15	147	80	19
Raters 1 & 3	105	90	10	126	87	13
Total	252	90	10	441	84	16

Notes: 15 test forms were randomly selected for multiple scoring. Each test form contained one teacher's responses to 21 items. Some of the test forms were scored by all three raters and thus are double counted in the *n*'s shown.

This combined measure showed higher reliability (coefficient alpha) of .84 and .89 for the total pre- and post-test scores, respectively. Nonetheless, the measurement error inherent in both scoring and scores suggests caution in the interpretation of these scores.

Teachers' assessment practices

Data on teachers' perceptions of their assessment practices were derived from weekly, on-line logs that teachers completed as part of the study. The logs, originally

Table 3. Score reliabilities (coefficient alpha) for teacher performance assessment scores and total composite score including multiple-choice content measures.

Assessment	Total score possible	n*	Pre-test α	*n*	Post-test α
Content knowledge	7	38	.51	28	.48
Analysis and interpretation	21	38	.73	28	.81
Next instructional steps	21	38	.79	28	.84
Total composite score	79**	34	.84	28	.89

Notes: **n*'s reflect all teachers completing an instrument and pre- and post-test *n*'s are not fully overlapping. Those who completed a post-test did not necessarily complete a pre-test.
**Combines teachers' scores on the three performance assessment components (49 total points) with scores from the multiple-choice content measure (30 total possible points) for a total of 79.

designed as a fidelity of implementation measure for the larger study, asked teachers to report on:

(1) how much time they spent teaching the curriculum;
(2) their use of available assessment tools and strategies (e.g. daily curriculum mapping of learning goals and activities, data sheets, student self-assessment, end of unit assessments);
(3) the time and strategies used to analyse student work (e.g. analysis of student notebooks, observation of inquiry); and
(4) their provision of feedback and next step strategies.

The log featured two types of questions. In responding to questions about time, teachers were asked to indicate the number of days during a week and the time per day that was spent teaching the curriculum and, if any, analysing student work. Questions about teachers' use of assessment strategies and tools asked teachers to indicate the number of days that week they had used each strategy. Because the number of days a strategy could be used was dependent on the number of days the curriculum was taught and because we wanted a comparable measure across teachers, we standardised the use questions to reflect the proportion of days taught that each strategy was used – this is, the number of days a strategy was used during a week was divided by the number of days the curriculum had been taught that week. Scores on each item were then summed across all completed logs for each teacher, and a mean score was created to represent each teacher's perception of assessment use.

Although all study teachers completed logs, the number of weekly logs completed varied greatly. Some teachers completed as few as two logs, while at the other extreme, other teachers completed more than 20 across the two curriculum modules they taught as part of the study. In fact, the sample's modal range was 20–24 logs completed, and the median response was in the range of 11–15 logs completed. Because teachers' mean log scores were used to estimate their assessment use, estimates were more reliable for some teachers than for others and this source of error remains unexamined.

Factor and conceptual analysis of the log data were used to distil assessment use variables for analysis. These analyses revealed one overriding factor, as shown in Table 4, that seemed to represent perceived intensity of on-going assessment use. That is, the factor encapsulated all of the log items dealing with on-going assessment strategies, combined with the amount of time spent each week on the

Table 4. Factor 1: Intensity of teacher reported on-going assessment use. Principal component analysis of mean teacher log responses (Varimax rotation, $n = 203$ teachers*).

Log item	Factor 1 loading
Frequency of use**	
Used daily curriculum mapping	.68
Planned and used assessment	.69
Analysed student work in notebook	.79
Analysed student work on response sheets	.72
Analysed observations of students	.74
Recorded and used assessment information on informal data chart	.65
Provided feedback to individual student based on analysis of student work	.67
Provided feedback to the entire class based on analysis of student work	.64
Retaught content	.61
Times curriculum taught per week	.76
Time spent per day analysing student work	.68

Notes: *This large n was available from another component of the larger study from which this study is derived and included 2327 logs.
**Frequency of use was computed as the number of days a week each strategy was reported relative to the total number of days the curriculum module was taught that week.

curriculum and on analysing student work. Teachers who were high on the items in this factor reported engaging more frequently in assessment practice, and their assessment and instruction were more concentrated (based on number of days taught) than those who were lower on the factor. Raw scores for items loading on this factor were converted into z scores, and z scores were summed to create a total perceived 'intensity of on-going assessment use' score for subsequent analysis. The scores exhibited high reliability (alpha = .95).

Although interpretation of log scores is limited by their self-report nature, scores did show an encouraging relationship to independent, observation measures of teachers' assessment practices. As part of the larger study, a subsample of teachers ($n = 12$) was observed about the same assessment practices queried by the log, e.g. teachers' inspection/analysis of student notebooks, response sheets and provision of feedback. Observation data were used to create an overall assessment implementation variable, denoting the number and quality of assessment practices in evidence. The Pearson correlation coefficient revealed a moderately strong relationship – .75 – between the observation and log (factor 1) measures (for more detail, see Osmundson et al., 2011).

Student achievement data

Student achievement measures included a specially developed end-of-year (EOY) assessment that addressed core topics within the three modules and the end-of-year standardised science assessment that was administered as part of the state assessment programme. Because the state test was designed to assess a total of 18 topics of which study modules reflect only a small sample, we were concerned about the state

assessment's sensitivity to the study intervention. Students' prior year (third grade) scores from the state reading assessment served as proxies for students' prior achievement and were used as covariates in study analyses, as were available data on student demographic characteristics. Reported internal consistency (coefficient alpha) of the state reading measure is .93 (Pearson, 2009), and for the fourth-grade end-of-year science assessment, reported internal consistency is .90 (Pearson, 2010).

The EOY assessment was specially developed by our research partner, WestEd, to address the content of the three modules that were part of the larger study: (1) magnetism and electricity, (2) water and (3) structures of life. Administered at the end of the study year, the assessment comprised 30 multiple-choice questions, 10 items on each of the three content areas. Test reliability was estimated at .76 (KR20) and standard error of measurement estimated at 2.57 (Timms et al., 2013) which contributes additional noise in detecting treatment or other effects. In addition, because all teachers in the study taught two modules, magnetism and electricity plus either water or structures of life, students had the opportunity to learn only two of the three areas assessed in the EOY assessment.

Analysis

Descriptive statistics, regression, hierarchical modelling 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 on teacher content knowledge. Study analyses used an alpha level of .05 for all statistical tests.

Results

Teacher knowledge: multiple-choice pre–post content survey results

The results presented in Table 5 show that for both groups of teachers, post-test scores on the multiple-choice content test were higher than pre-test ones prior to the start of the study. Given that test items were designed for elementary school students, it is surprising that teachers averaged only about 67% (20/30) correct on the pre-test. However, after teaching the curriculum for two years, both groups substantially increased their scores. While treatment teachers showed higher gains pre–post, we did not find statistically significant differences in teachers' post-test knowledge on the multiple-choice test as a function of treatment condition, although the small sample size with complete data-sets (n = 24 total teachers) severely limits statistical power.

Teacher knowledge: content-pedagogical performance assessment results

Table 5 also shows teachers' pre- and post-scores on each component of the content-pedagogical performance assessment. For both groups, initial scores on the content component were quite modest, similar to results on the multiple-choice content test. Scores on the remaining two performance components – analysis and interpretation of student work and knowledge of instructional next steps – were substantially lower. Teachers' mean scores ranged from 29% to 37% correct of the total points possible. While treatment teachers' scores appear slightly higher than those of the

Table 5. Mean teacher pre/post content-pedagogical knowledge scores based on matched pre–post scores available for each measure.

Assessment	Total possible score	Control		Treatment		Group difference treatment-control	
		Pre (SE)	Post (SE)	Pre (SE)	Post (SE)	Pre-test (SE)	Post-test (SE)
Content multiple-choice ($n = 24$, 11 control, 13 treatment)	30	20.00 (.63)	25.18 (.72)	19.92 (.80)	26.69 (.44)	−.08 (1.04)	1.51 (.82)
Performance assessment ($n = 28$, 13 control, 15 treatment):							
• Content	7	3.64 (.45)	5.18 (.42)	4.15 (.40)	6.54 (.14)	.52 (.61)	1.36* (.42)
• Analysis/interpretation	21	5.63 (.95)	10.72 (.90)	7.46 (.76)	14.00 (.75)	1.83 (1.20)	3.27* (1.16)
• Instructional next steps	21	5.55 (.96)	11.45 (1.00)	6.77 (.85)	14.00 (.85)	1.22 (1.27)	2.55 (1.31)
Total score ($n = 24$, 11 control, 13 treatment)	79	34.82 (1.86)	52.55 (2.51)	38.31 (2.40)	61.23 (1.96)	3.49 (3.12)	8.69* (3.14)

*Statistically significant at $p = .05$.

control teachers at the beginning of the study, the differences are not statistically significant.

Similar to the multiple-choice trends, scores on content-pedagogical performance assessment also improved for both groups after teaching the study curriculum for two consecutive years. The descriptive results suggest that both groups showed the largest gains in the area of instructional next steps, and that treatment group gains appear larger than those for control group.

Regression analyses were conducted to test the statistical significance of treatment effects on teachers' content-pedagogical knowledge. Using pre-test scores for each teacher performance assessment component as a covariate, results showed statistically significant treatment effects. For all three areas – content, analysis and interpretation and instructional next steps – treatment teachers outperformed control teachers. Table 6 displays results of the regression analyses. While the sample size and attrition limit generalisability, these differences provide evidence of the impact of the treatment on participating teachers' content-pedagogical knowledge.

Assessment implementation: log results

Table 7 summarises the descriptive results from the logs, using the teacher as the unit of analysis and teacher mean scores for each log item over the course of the curriculum unit. On average, the results reported in Table 7 indicate that all teachers used the study modules approximately three times a week and, on each of these days, reported spending 5–10 minutes analysing students' work on the modules. While average responses were generally similar between treatment and control teachers, it is noteworthy that treatment teachers spent substantially more time looking at student work and were more likely to use the curriculum maps provided with the modules that provided teachers with an overview to daily learning objectives and activities. Teachers reported that, of the options provided, they most frequently used observation to gauge how students were doing. Teacher log responses indicated that teachers engaged in such observation on about half the days they taught the modules. However, looking across the categories, teachers on average reported

Table 6. Regression analysis of treatment effect: teacher content-pedagogical knowledge performance (n = 28 teachers, 13 control, 15 treatment).

Variable	df	Parameter estimate	Standard error	t value	Pr > |t|
Post-content (N = 28)					
Intercept	1	5.15	.58	8.96	<.0001
Pre-content	1	.06	.12	0.46	.65
Treatment	1	1.20	.39	3.06	.01
Post-analysis and interpretation (N = 28)					
Intercept	1	8.77	1.44	6.11	<.0001
Pre-analysis and interpretation	1	.31	.18	1.73	.10
Treatment	1	3.50	1.20	2.92	.01
Post-next step (N = 28)					
Intercept	1	9.16	1.46	6.27	<.0001
Pre-next step	1	.37	.19	1.98	.06
Treatment	1	2.73	1.31	2.08	.05

Table 7. Teacher log data descriptive results by group.

Teacher log questions	Control $n = 14$ teachers (SD) (SE)	Treatment $n = 16$ teachers (SD) (SE)	Difference (SE)
Number of times taught/week	2.9 (.91) (.24)	3.1 (.73) (.18)	.19 (.30)
Average (minutes/day analysing student work)	5.91 (5.8) (1.55)	10.77 (6.85) (1.71)	4.86* (2.34)
Frequency of assessment use**:			
Used daily goal mapping	.57 (.48) (.13)	.77 (.38) (.09)	.20 (.16)
Planned and used assessment	.56 (.23) (.06)	.52 (.21) (.05)	−.04 (.08)
Analysed student work in notebook	.35 (.33) (.09)	.43 (.23) (.06)	.08 (.10)
Analysed student work on response sheets	.46 (.28) (.08)	.33 (.22) (.06)	−.12 (.09)
Analysed observations of students	.52 (.32) (.09)	.48 (.24) (.06)	−.04 (.10)
Recorded and used assessment information on informal data chart	.15 (.19) (.05)	.15 (.18) (.05)	.002 (.07)
Provided feedback to individual student based on analysis of student work	.4 (.23) (.06)	.37 (.22) (.05)	−.03 (.08)
Provided feedback to the entire class based on analysis of student work	.58 (.27) (.07)	.41 (.18) (.05)	−.17 (.08)
Retaught content	.23 (.26) (.07)	.15 (.08) (.02)	−.07 (.07)

Notes: *Statistically significant at $p = .05$.
**Scale = percentage of days taught that assessment strategy or tool was reported in use.

analysing student work in science notebooks or response sheets or observing students' work on at least a daily basis. Similarly, while teachers report providing feedback to individual students or to the entire class the majority of the time, their self-reports indicate that they did not consistently provide students' feedback based on their analysis of student work – i.e. they more frequently reported some form of analysis than providing feedback. Formal recording of data, using the data charts supplied with the curriculum, was the least frequently used strategy, followed closely by reteaching content.

Table 8 displays the 'intensity of on-going assessment use' scores for control and treatment teachers, which were based on the log factor analysis described above. Results indicate that treatment teachers engaged in significantly more assessment than did control teachers, as judged by a composite of weekly time spent in the curriculum, time spent analysing student work and use of various assessment strategies. These findings not only suggest a treatment effect but also support our plan to use the treatment condition as a proxy for assessment use.

Table 8. Perceived intensity of on-going assessment use estimated from principal component factor scores, using standardised variables scores.

Group	*n*	Intensity of on-going assessment (Mean)	SD	SE
Control	14	−.82	1.07	.29
Treatment	16	−.07	.88	.22
Difference		.75*	.98	.36

*Statistically significant at $p = .05$.

Student outcomes

Table 9 summarises students' scores on measures used to evaluate student learning effects. Students' third grade scores on the state reading assessment for the year prior to the study provided a gauge of students' entering achievement levels. Mean scale scores for treatment and control groups were 470 and 466, respectively, which the state classifies as 'meet state standards' for grade-level proficiency. Because the state classifies students with scores in the range of 431–515 as 'meet state standards' for grade-level proficiency, both groups scored near the middle of this range, which was slightly above the state-wide mean score of 456 (Arizona Department of Education, 2009). Although treatment students' entering scores were slightly higher than those of the control students, differences are not statistically significant. These data indicate that the two groups were comparable in achievement prior to the start of the study.

Results from the specially developed end-of-year (EOY) science assessment indicate that, on average, students achieved about 60% correct. Treatment and control groups' scale scores from the end-of-year state science assessment were 522 and 533 for the control and treatment groups, respectively. As with the state reading scores, these mean scores fall within the state's classification of 'meet state standards' for proficiency, which the state defines as scores of 500–546 (Arizona Department of Education, 2010b). Note that mean scores for both groups are considerably above the state mean of 433 (Arizona Department of Education, 2010a). Scores on both the EOY assessment and the state science assessment show a small advantage for treatment students relative to control students, both of which are statistically significant based on the descriptive results alone.

Because of the complex nature of the study data, where students are nested within classrooms and classrooms within schools, we used Hierarchical Linear Modelling (HLM) to further examine group differences in the end-of-year science measures. The HLM analysis was conducted in a step-wise manner. First, we entered the treatment variable and students' prior achievement scores (state reading assessment scores before the study began) to control for potential differences in students' entering achievement levels. In the second step, we added additional student

Table 9. Student achievement scores: entering ability and end-of-study science assessment scores for treatment and control classrooms (n = 30 teachers, 14 control, 16 treatment).

	Treatment students			Control students			
Test	n	Mean (SE)	SD	n	Mean (SE)	SD	Difference (SE)
Prior year state reading assessment	429	470.70 (2.48)	51.32	355	465.8 (2.78)	51.95	4.91 (3.71)
End-of-year science assessment	438	19.27 (0.21)	4.45	365	18.21 (.26)	5.04	1.07* (.33)
State assessment: science	448	533.0 (2.56)	54.47	370	521.5 (2.64)	51.82	11.47* (3.70)

*Statistically significant at p = 0.05.

background variables (e.g. gender, ethnicity, language status), and in the final step, added the interaction terms between the treatment variable and student background variables.

Table 10 shows the results of the final model. These results indicate that students in treatment classrooms outperformed those in control classrooms close to statistical significance (p-value = .053). The estimated treatment effect is 1.49, which indicates that, holding student background characteristics and prior ability constant, a student from the treatment group will have an EOY score 1.49 points higher than his or her peers from the control group. While these findings are statistically and practically weak, they should be interpreted in the context of the reliability and precision of the available outcome measure as well as imperfect alignment between the measure and students' opportunity to learn. That is, as mentioned above, students in the study had the opportunity to learn only two of the three topics addressed by the end of year measure: magnetism and electricity plus either water or structure of life.

As would be anticipated, students' entering ability, as measured by the prior year state reading scores, had a statistically significant impact on end-of-year scores, as did students' language status. English language learners scored lower than fully English-proficient students. No significant interactions with treatment were observed, beyond a practically trivial difference in initial reading scores (and which were controlled, as part of the analysis). A similar model was tested for the results of the state science assessment, but no significant effect of treatment was found.

Table 10. Treatment effects on end-of-year science assessment (HLM fixed effects model).

Fixed effects	Coefficient	SE	df	t-value	p-value
Model for class mean					
Intercept, $\beta 00$	18.38	.77	27	23.85	<.0001
Treatment effect (1-treatment, 0-control), $\beta 01$	1.49	.74	27	2.03	.053
Third grade scale score (reading), $\beta 02$	.01	.01	27	.49	.62
Third grade ethnicity: Hispanic, $\beta 03$	−.74	.61	587	−1.21	.23
Third grade ethnicity: white, $\beta 04$	1.07	.56	587	1.90	.06
Third grade English language learner (1-yes, 0-no), $\beta 05$	−1.75	.81	587	−2.15	.03
Third grade free/reduced lunch (1-yes, 0-no), $\beta 06$	−.64	.55	587	−1.16	.25
Third grade gender (1-male, 0-female), $\beta 07$	.87	.49	587	1.78	.08
Treatment/control interaction with third grade ELL, $\beta 08$	−.46	1.07	587	−.43	.67
Treatment/control and third grade free/reduced lunch interaction, $\beta 09$	−.96	.74	587	−1.29	.20
Model for third grade scale score (reading) slope					
Intercept, $\beta 10$	.03	.00	587	6.20	<.0001
Treatment, $\beta 11$	−.01	.01	587	−1.97	.05

Notes: Level 1, n = 630 students; level 2, n = 30 teachers, 14 control, 16 treatment.

Path analysis results

We conducted teacher-level path analysis to investigate relationships among teachers' content knowledge, assessment use, student prior reading scores and student post-scores. Students' test scores were aggregated to teacher levels to indicate the average student's performance for each teacher. The path model controlled for students' entering achievement level and predicted those teachers' assessment practices, as measured by self-reported intensity of assessment use and student learning would be directly related to condition. Teachers' content knowledge, however, 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 sophisticated use of assessment, and as a result of that engagement, greater intensity of assessment was expected to be positively related to student learning. However, as discussed earlier, we viewed teachers' assessment practices as a key intervening variable between teacher knowledge and student learning, and 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' initial content-pedagogical knowledge was used. The composite measure combined results across the multiple-choice and performance assessment measures administered at the start of the study. The self-report log measure of intensity of on-going assessment use represented teachers' use of assessment. Models were tested using both the specially developed end-of-year student assessment and the state science assessment results as the indicator of student learning. All models were tested at the teacher level and used students' prior year (third grade) state reading scores to control for any differences in entering student ability.

Moreover, to examine the extent to which teachers' use of assessment might positively influence their content-pedagogical knowledge, analyses also examined the relationship between teachers' intensity of on-going use of assessment and both teachers' content-pedagogical knowledge and changes in teachers' content-pedagogical knowledge. We also tested our assumptions of the indirect vs. direct effect of teacher content-pedagogical knowledge on student learning by running multiple models (i.e. the model including a direct effect only, an indirect effect only and both direct and indirect effects).

Figure 2 displays the results of the path analysis for the specially developed end-of-year science measure, which was tailored to module content. (As expected, the broader state science assessment was not sensitive to study variables.) The standardised path coefficients shown in Figure 2 indicate how much a change of one standard deviation in a prior variable would produce on the subsequent variable (also in standard deviation units).

Results show that students' prior achievement, as measured by state reading scores, has a strong relationship to student science outcomes, much stronger than any other variable in the study. After controlling for prior achievement, the intensity of teachers' self-reported on-going 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 .30 standard deviation units in students' performance. As hypothesised, teachers' content-pedagogical knowledge has no direct relationship to student learning, but shows a small but

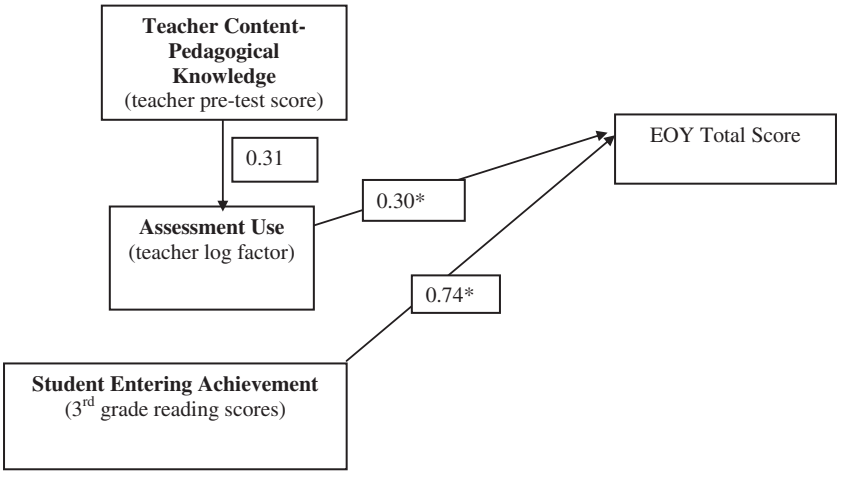

Figure 2. Standardised path coefficient model (n = 25 teachers).
*Indicates statistically significant at p = 0.05.

non-statistically significant relationship with teachers' assessment use. Bentler's Comparative Fit Index shows a good fit for the model displayed (CFI = .996, RMSEA = .046).

Study analyses found no relationship between intensity of self-reported assessment use and teachers' content-pedagogical knowledge. That is, teachers who reported 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 engaged in those tasks.

Discussion and conclusion

This paper started with three research questions. We end it by summarising our findings with regard to each question and then consider implications. In doing so, we consider the limitations and challenges of this study: a small sample size, with attrition, that limits both the power of the study to detect treatment effects and study generalisability; the challenges of measuring the complex interaction of teacher content and assessment knowledge and the implications of the reliability of study measures, which added noise to study analyses; and standard errors of measurement that give further pause about the significance of observed differences. These limitations certainly suggest caution in drawing any firm conclusions about study questions.

What is the quality of teachers' content-pedagogical and assessment knowledge?

The study used multiple-choice and performance assessments to measure teachers' content and pedagogical knowledge. The multiple-choice test was drawn from publicly available items on magnetism, electricity and electromagnetism intended for elementary students. By drawing on items intended for students, the multiple-choice measure set an intentionally low bar for teachers' content knowledge, and the reliability of the multiple-choice items was less than optimal. Nonetheless, at the

beginning of the study, both groups of teachers only answered approximately 60% of the items correctly. By the end of the study, after having taught the study curriculum twice, performance ranged from 85% to 92% correct for the control and treatment groups, respectively. Treatment teachers showed higher pre-to-post gains but differences were not statistically significant.

The results of the teacher content-pedagogical performance assessment showed these same general trends in both the reliability of the scores and the descriptive findings. The measure asked teachers to respond to seven sets of items dealing with magnetism, electricity and electromagnetism. In each set, teachers first were asked to respond to a constructed response problem that required them to explain a relevant concept and then to analyse and interpret student responses to that same problem and then to use their analysis to specify next steps in instruction to respond to students' learning needs. Subject to both rater and score variation, the performance assessment scores evidenced considerable error.

Descriptive statistics showed that teachers' pre-test performance was low in all three areas at the start of the study and improved on the post-test, at the end of the study. Despite improvement, however, teachers' performance at the end of the study remained relatively low. Performance was particularly low for tasks involving the analysis and interpretation of student work and those on formulating next steps for instruction. 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. in analysing student responses and identifying implications for subsequent instruction, also bears directly on participating teachers' assessment capacity and the potential value of their using assessment to improve student learning. Despite study limitations of sample size and measure reliability, these results are consistent with prior studies documenting the difficulty teachers experience in two critical aspects of formative assessment practice (Herman et al., 2010; Heritage et al., 2009). If teachers' analysis of student work does not result in accurate diagnosis of student learning needs, teachers' use of assessment may add error rather than knowledge to instructional planning and decision-making. Furthermore, if teachers cannot use their inferences to form effective next steps, assessment's value for improving learning is sharply reduced.

What is the relationship between teacher knowledge and assessment practice?

We expected a positive relationship between teachers' content-pedagogical knowledge and their assessment practices because we supposed that if teachers spent more time engaged in assessment and analysing student work, they would gain insight on how students' knowledge develops and the misconceptions and obstacles that may occur along the way. Because the study treatment involved systematically embedding formative and end-of-investigation assessments in a hands-on science curriculum and encouraged teachers to use assessment results, observed differences between treatment and control groups provided one test of this hypothesis. Aided by curriculum-embedded assessments that systematically elicited and diagnosed students' understanding of key ideas and supported by scoring and interpretation guides for analysing student work and planning next steps, teachers in the treatment group, we thought, should exhibit greater gains in content-pedagogical knowledge than

would control teachers. The relative gains in content-pedagogical knowledge for treatment and control teachers supported this hypothesis. Controlling for teachers' pre-study knowledge, regression analyses of teacher's content-pedagogical knowledge at the end of the study revealed a statistically significant treatment effect. For all aspects of content-pedagogical knowledge assessed – content, analysis and interpretation and instructional next steps – our small sample of treatment teachers outperformed control teachers.

However, our direct analysis of the relationship between teachers' assessment use and teacher knowledge showed no evidence of relationship. We hypothesised a significant, positive relationship between the two, based on the rationale above. However, our path analyses examining the relationship between the intensity of teachers' assessment use and changes in teachers' content knowledge revealed no measurable connection. Higher intensity of assessment use, as measured by teachers' responses to weekly logs, was not associated with improvement in teachers' content-pedagogical knowledge. Nor was there any relationship between teachers' use of assessment and their content-pedagogical knowledge at the end of the study.

Descriptive findings from the log, however, may provide a clue to the lack of relationship. Log results indicated that teachers did not consistently analyse student work in depth. On average, teachers reported spending only 5–10 minutes analysing students' written responses or other forms of work, and thus, the intensity of our teacher sample's assessment use may have been insufficient to enable a measurable effect. In addition, the fact that only the treatment teachers had the benefit of established, high-quality assessments may have diminished the effect of assessment use across the entire sample. Assessments that do not well elicit student thinking, that do not focus on the most important aspects of a domain and/or are not accurately analysed, may have limited value in supporting teacher knowledge development. A within-treatment analysis of the relationship between assessment use and teacher knowledge could have helped to clarify this issue. However, the sample size was too small for such an analysis.

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

The results from the study's hierarchical linear modelling (HLM) and path analysis results lend evidence on the effect of formative assessment on student learning and the paper's hypothesis about the relationship between teachers' use of assessment in instruction and student learning. HLM results, controlling for students' entering reading ability and accounting for the nested nature of the data, suggested that students in treatment classrooms scored slightly higher (at a marginally statistically significant level, $p = .053$) than those in control classrooms on the specially developed, end-of-year science study measure. Trends were similar on the end-of-year state science assessment, but differences were not statistically significant. We interpret these differences in performance as an effect of the use of the embedded assessment system, since the inclusion of new assessment and assessment guidance was the primary difference between treatment and control conditions. Results from teacher logs support this interpretation by demonstrating that assessment use was more intensive in treatment relative to control classrooms.

Our path analysis further explored the relationship between assessment practice and student learning by considering teacher content and pedagogical knowledge as an intervening variable. We hypothesised that teachers' content-pedagogical knowledge should have a direct effect on their assessment use, but would not directly affect student learning. That is, teachers with more sophisticated content-pedagogical knowledge would be more likely to engage in on-going assessment and, further, the application of this more sophisticated knowledge in the process of assessment to inform instruction would increase student learning. Teacher knowledge thus was expected to have an indirect effect on student learning through the mechanism of teachers' assessment practices. Path analysis results provided some support for our hypotheses: controlling for students' entering achievement, as measured by standardised test scores in reading, teachers' use of assessment, as measured by a composite of teachers' responses to weekly logs, was positively related to student learning outcomes. Greater self-reported intensity of assessment use during instruction was associated with *more* science learning at the end of the study. However, the study found no evidence of a direct, statistically significant relationship between teachers' initial content-pedagogical knowledge, as gauged by multiple measures administered at the start of the study, and student learning. Results showed a small positive relationship between teachers' initial content-pedagogical knowledge and teachers' assessment use, but the relationship was not statistically significant and at a level of little practical significance. The overall model fit was high for these analyses, which provides some level of confidence in the findings here, but the practical magnitude of the observed relationships is relatively small.

Conclusions

Study findings, despite limitations of sample size and measure reliability and the challenges in measuring the complex cognitive process of assessing student work, add to the evidence base documenting the positive effects of formative assessment, or at least one important element of it: students whose teachers spend more time and who more frequently engage in analysing and providing feedback on student work, as gauged by the overall log factor scores, achieve higher learning than do 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 discussion or through analyses of student responses in science notebooks and other written responses – is associated with higher student performance.

Although observed relationships need replication, they are noteworthy in the light of both the weaknesses in the reliability of the teacher measures and the low initial content and pedagogical knowledge they show for study teachers. These scores revealed disappointing levels of expertise, particularly with regard to teachers' ability to accurately interpret student work and to use their analysis to figure out appropriate next steps. Yet, absent such expertise, teachers will have difficulty accurately analysing the gap between where students are and where we want them to be relative to important learning goals, or taking effective action to close the gap (see also Heritage, 2012; Sadler, 1989), and the learning benefits of using assessment to inform instruction likely will be curtailed.

Both treatment and control teachers showed pre- to post-study gains in content and pedagogical knowledge. This improvement may be due at least partially to a testing effect as the same assessment was given pre-intervention and two years later,

at the end of the study. However, since the treatment has a measurable effect on teachers' content-pedagogical knowledge, the study provides some support for the hypothesis that use of validated assessments that well address student learning progress and elicit student thinking can contribute to the development of stronger teacher knowledge. The assessments may serve to focus teachers' attention on core goals, how learning is expected to develop and obstacles and misconceptions that may occur along the way. Admittedly, the path model examining this connection shows only a tenuous connection, and analyses between assessment use and changes in teacher knowledge failed to show any relationship. However, the small sample size, particularly of the treatment group, was a clear constraint for such inquiry. Further, log findings suggested that sampled teachers did not consistently engage in ongoing assessment of student learning. Teachers' analysis of student work did not consistently occur; and other strategies also appeared sporadic – or at least less than everyday routine. Higher benefits may have accrued with higher engagement with assessment.

In this and other study limitations, we see important avenues for future research. The quality of measures of teachers' content-pedagogical knowledge and of assessment use was an impediment for this study and also presents a challenge for research on formative assessment. The field would benefit from the development and validation of more robust measures. Further, the level and variation in teacher content-pedagogical knowledge among study participants likely constrained study findings. Future studies should investigate hypothesised relationships with samples showing larger variance, including teachers with higher levels of content-pedagogical knowledge as well as larger, representative samples to support the generalisability of study findings. For example, future studies might provide stronger mechanisms to assure that teachers had the content and assessment practice they needed, perhaps through more extended professional development and more guided practice analysing and interpreting the work than provided by this study. In this study, designed as an efficacy study rather than one focused on teacher professional development to build science and assessment knowledge, teachers participated in only three professional development days, which was likely insufficient to build their adequate content and pedagogical knowledge.

We conclude with the same issue we used in titling the paper, that is, the dynamic nature of formative assessment. Study findings suggest that teachers' use of formative assessment can benefit student learning. Yet, effective formative assessment practice places heavy demands on teachers' content and pedagogical knowledge, knowledge that may be spotty and incomplete based on the results of this study and others' research. Can analysing student responses to curriculum-embedded tasks help teachers strengthen their content and pedagogical knowledge, or is some minimal level of content knowledge necessary before teachers can effectively use assessment to benefit student learning? What are optimal approaches for developing teachers' capacity in these areas? Do teachers learn content more effectively in the process of studying content or in the process of analysing students' learning? This study raises possibilities, but the dynamics of these relationships and the hierarchies of influence remain unresolved.

The popularity of formative assessment in current policy initiatives provides opportunities to further investigate these dynamics. Study results underscore both the potential for and the challenge of bringing these policy initiatives to fruition, as well as the need for additional research.

Funding

The work reported herein was supported by prime sponsor number R305B070354 from the US Department of Education to WestEd, grant #R305B070354. 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, CRESST, or the US Department of Education.

Notes

1. We recognise that the term pedagogical-content knowledge (PCK) is conventionally used in the literature, but to differentiate our conception and measure, we use the term content-pedagogical knowledge throughout the manuscript.
2. M. Tiu, personal communication to author, 2010.

Notes on contributors

Joan Herman is Director Emeritus and current senior research scientist at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA. Her research explores the effects of accountability and assessment on schools and teachers, and the design of assessment systems to support educational improvement. Her recent work focuses on the quality and consequences of teachers' formative assessment practices and the assessment of twenty-first-century skills. She also has wide experience as an evaluator of school reform.

Ellen Osmundson is project coordinator of the University of California Office of the President's Innovative Learning Technology Initiative (ILTI). She was a senior research associate at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA when the study was completed.

Yunyun Dai was a senior research associate at the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) at UCLA for all study analyses. She currently serves as Statistician for the University of Southern California's Office of Admission and Planning.

Cathy Ringstaff is a senior research associate at WestEd. Her work focuses on the evaluation of mathematics, science and technology programmes.

Michael Timms is director of assessment and psychometric research at the Australian Council for Educational Research (ACER). He was the project director of the reported study and was at WestEd at the time it was completed.

References

Arizona Department of Education. (2009). *AIMS scale scores and performance levels*. Retrieved from http://www.azed.gov/research-evaluation/files/2011/08/spring2009scalescoretable.pdf

Arizona Department of Education. (2010a). *AIMS results 2010*. Retrieved from http://www.azed.gov/research-evaluation/aims-assessment-results/

Arizona Department of Education. (2010b). *AIMS scale scores and performance levels*. Retrieved from http://www.azed.gov/research-evaluation/files/2011/08/2010scalescoretable.pdf

Ball, D. L., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of Teacher Education, 59*, 389–407.

Bennett, R. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*, 5–25.

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

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*, 7–74.
Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*, 139–148.
Black, P., & 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: 103rd Yearbook of the National Society for the Study of Education* (pp. 20–50). Chicago, IL: University of Chicago Press.
Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability, 21*, 5–31.
Council of Chief State School Officers (CCSSO). (2008). *Attributes of effective formative assessment*. Washington, DC: CCSSO, FAST-SCASS.
Fennema, E., & Franke, M. (1992). Teachers' knowledge and its impact. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 147–164). New York, NY: Macmillan.
Formative Assessment for Students and Teachers (FAST), State Collaborative on Assessment and Student Standards (SCASS). (2008, October). *Attributes of effective formative assessment*. Washington, DC: Council of Chief State School Officers.
Furtak, E., Ruiz-Primo, M., Shemwell, J., Ayala, C., Brandon, P., & Shavelson, R. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education, 21*, 360–389.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.
Heritage, M. (2010, October). *Formative assessment and next-generation assessment systems: Are we losing an opportunity?* Report prepared for the Council of Chief State School Officers, Washington, DC.
Heritage, M. (2012). *Formative assessment in practice*. Boston, MA: Harvard Education Press.
Heritage, M., Jones, B., & White, E. (2010, April). *Supporting teachers' use of formative assessment evidence to plan the next instructional steps*. Paper presented for a symposium entitled Assessment: Evidence in Action, American Educational Research Association (AERA) conference, Denver, CO.
Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice, 28*, 24–31.
Herman, J., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2006). *The nature and impact of teachers' formative assessment practices* (CRESST Report No. 703). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
Herman, J., Osmundson, E., & Silver, D. (2010). *Capturing quality in formative assessment practice: Measurement challenges* (CRESST Report No. 770). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42*, 371–406.
James, M., Black, P., Carmichael, P., Drummond, M.-J., Fox, A., & MacBeath, J. (2007). *Improving learning how to learn in classrooms, schools and networks*. London: Routledge.
Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*, 28–37.
Little, J. W. (2003). Inside teacher community: Representations of classroom practice. *Teachers College Record, 105*, 913–945.

Long, K., & Malone, L. (2006, Spring). *Assessing science knowledge (The ASK Project), FOSS Newsletter #27*. Berkeley, CA: Lawrence Hall of Science. Retrieved January 28, 2015, from http://lhsfoss.org/newsletters/archive/FOSS27.assessing.html

Organisation for Economic Co-operation and Development (OECD). (2005). *Formative assessment: Improving learning in secondary classrooms*. Paris: OECD Publishing.

Osmundson, E., Herman, J., & Dai, Y. (2011). *Year 3 ASK/FOSS efficacy study* (CRESST Report No. 782). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Pearson. (2009). *2009 AIMS technical report*. Retrieved from http://www.azed.gov/standards-development-assessment/files/2011/12/aimstechreport2009.pdf

Pearson. (2010). *2010 AIMS technical report*. Retreived from http://www.azed.gov/standards-development-assessment/files/2011/12/aimstechreport2010.pdf

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

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

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–22.

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

Timms, M., Ringstaff, C., Schneider, S., Huang, K., Li, L., Tiu, M., … Dai, Y. (2013). *ASK/FOSS efficacy study: Final report*. San Francisco, CA: WestEd.

US Department of Education. (2010, September). *U.S. Secretary of Education Duncan announces winners of competition to improve student assessments*. Retrieved from http://www.ed.gov/news/press-releases/us-secretary-education-duncan-announces-winners-competition-improve-student-asse

Wilkins, J. L. (2008). The relationship among elementary teachers' content knowledge, attitudes, beliefs, and practices. *Journal of Mathematics Teacher Education, 11*, 139–164. doi:10.1007/s10857-007-9068-2

Wilson, M., & Draney, K. (2004). Some links between large-scale and classroom assessments: The case of the BEAR assessment system. *Yearbook of the National Society for the Study of Education, 103*, 132–154.

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