Instr Sci (2016) 44:267–291
DOI 10.1007/s11251-016-9371-3

Teachers' formative assessment abilities and their relationship to student learning: findings from a four-year intervention study

Erin Marie Furtak[1] · Katharina Kiemer[2] · Ruhan Kizil Circi[1] · Rebecca Swanson[1] · Vanessa de León[1] · Deb Morrison[1] · Sara C. Heredia[3]

Received: 14 February 2015 / Accepted: 9 April 2016 / Published online: 16 April 2016

Abstract The teaching practices of recognizing and responding to students' ideas during instruction are often called formative assessment, and can be conceptualized by four abilities: designing formative assessment tasks, asking questions to elicit student thinking, interpreting student ideas, and providing feedback that moves student thinking forward. While these practices have been linked to positive learning outcomes for students, designing and enacting formative assessment tasks in science classrooms presents instructional challenges for teachers. This paper reports on the results of a long-term study of high school biology teachers who participated in a 3 year professional development program, called the Formative Assessment Design Cycle (FADC), which guided them to iteratively design, enact, and reflect upon formative assessments for natural selection in school-based teacher learning communities. Nine teachers participated for three academic years; sources of data included teachers' interpreting of student ideas in line with a learning progression, the formative assessment tasks they designed each year of the study, videotaped classroom enactment of those tasks, and pre-post test student achievement from the Baseline and final year of the study. Results indicate that, on average, teachers increased on all abilities during the study and changes were statistically significant for interpreting students ideas, eliciting questions, and feedback. HLM models showed that while only the quality of feedback was a significant predictor at Baseline, it was teachers' task design and interpretation of ideas in Year 3. These results suggest the efficacy of the FADC in supporting teachers' formative assessment abilities. Findings are interpreted in light of professional development and formative assessment literatures.

Keywords Science teaching · Professional development · Student achievement · Formative assessment · Secondary science · Biology education

✉ Erin Marie Furtak
erin.furtak@colorado.edu

[1] School of Education, University of Colorado at Boulder, UCB 249, Boulder, CO 80309, USA

[2] TUM School of Education, Munich, Germany

[3] The Exploratorium, San Francisco, USA

Introduction

In the US, the teaching practices of recognizing and responding to students' ideas in the course of instruction are often called formative assessment (National Research Council [NRC] 2001a). These practices entail teachers creating opportunities for students to share their ideas as they develop during instruction, identifying those ideas as they are shared, and providing feedback to move students forward in their learning (Shepard 2000). In science education, formative assessment is often described as the instructional tasks teachers enact to surface student thinking (Ayala et al. 2008), as well as the whole-class discussions teachers orchestrate as opportunities to attend and respond to students' ideas (e.g. Bell and Cowie 1999; Coffey et al. 2011; Duschl and Gitomer 1997). The benefits of formative assessment to support student learning have been established through several synthesis studies (Kingston and Nash 2011). Black and Wiliam (1998) noted that it was particularly successful at narrowing achievement gaps between low and high performing students.

Designing and enacting formative assessment tasks in science classrooms presents multiple challenges for teachers. Teachers must be able to design instructional experiences that will create opportunities for students to share their thinking, and then be able to navigate the types of ideas that are likely to come up. That is, teachers must be able to identify and interpret student ideas, as well as to design tasks and orchestrate classroom conversations so that student ideas might be shared (Furtak 2011). Learning progressions, or representations of how student ideas develop in conceptual domains, may support the development of these formative assessment abilities for teachers (Bennett 2011; Heritage et al. 2009). Little research has been completed that explores how to support teachers through professional development in designing their own formative assessment activities (e.g. Atkin et al. 2005); furthermore, the field is only beginning to examine how learning progressions can support teachers' formative assessment abilities (Alonzo and Gotwals 2012).

In this paper, we explore the ways in which teachers who participated in a 3-year professional development intervention changed in their abilities to design formative assessment tasks, explore student thinking through questions and feedback, and interpret this thinking. Furthermore, we explored the relationship between these formative assessment abilities and student learning.

Formative assessment

The phrase 'formative assessment' refers not only to the *activities or tasks* teachers use to create opportunities for students to share their thinking, but also the *instructional practices* of students and teachers as ideas are made explicit, and feedback is provided to advance student learning (Bennett 2011). In this paper, we build upon this distinction between formative assessment tasks and practices and define the construct of formative assessment as consisting of a set of four complementary abilities for teachers: designing tasks, asking questions to elicit student ideas, interpreting those ideas, and providing informational and constructive feedback to move thinking forward.

Designing formative assessment tasks

Traditional science classroom activities involve teacher lectures, recipe-style laboratories, and assessments that leave little space for students to develop and share their thinking (NRC 2001b). In contrast, formative assessment tasks are written activities that create opportunities for students to share their ideas (Kang et al. 2014). Formative assessment tasks are designed such that students have opportunities to explain their thinking (Cowie and Bell 1999) in response to written prompts with varying formats, such as open-ended, constructed-response questions (Ayala et al. 2008), multiple-choice plus justification questions (Furtak 2009a, b), and predict-observe-explain activities (White and Gunstone 1992).

Asking questions to elicit student ideas

Studies of science classrooms indicate that teachers still control the majority of classroom interactions (e.g. Jurik et al. 2013; Kobarg and Seidel 2007). In these traditional settings, classroom discourse is often constrained and evaluative with teachers asking simple questions and providing evaluative feedback (Mehan 1979; Cazden 2001), scarcely leaving time or space for students to voice their ideas and expand on their thinking (Seidel et al. 2007). Yet research into classroom discourse has shown that teachers asking open-ended, authentic questions (e.g. those starting with "Why," "How," or "What do you think?") can provide room for students to share their ideas (e.g. Cazden 2001; Michaels et al. 2008). Formative assessment classroom practices nurture a free exchange of ideas in which teachers encourage extended student contributions that contain substantive information about student thinking (Coffey et al. 2011).

Interpreting student ideas

Many science teachers view student ideas from a binary, "get it or don't" (Otero and Nathan 2008) perspective; however, research has indicated that student thinking is multifaceted, context-dependent, and develops over time (e.g. Smith et al. 1993). The purpose of formative assessment is to surface the true nature of student thinking so that teachers can listen to and build upon those ideas in order to inform their instruction; as such, viewing student thinking as complex is essential to supporting student learning (Furtak 2011).

Table 1 Four teacher formative assessment abilities

Ability	Traditional/authoritative	Formative assessment
Designing formative assessment tasks	Constrained tasks, few opportunities to elicit student thinking	Open-ended tasks designed to elicit student thinking
Asking questions to elicit student ideas	Closed-ended questions	Asking authentic, open-ended questions to surface student thinking
Interpreting student ideas	Right or wrong	Complex and spanning a continuum of understanding
Providing feedback	Evaluative responses	Pushing students to expand on their thinking and providing information to improve performance

Providing feedback

In traditional instruction, the teacher is viewed as the ultimate source of knowledge (‘primary knower’; Bernstein 2000), and student ideas are only drawn out for the purpose of evaluating them (Reznitskaya 2012; Mercer 2010; Alexander 2008). In formative assessment, however, teachers build on student ideas and provide helpful feedback to move students forward in their learning (Shepard 2000). In doing so, they provide information about the quality of student performance, cuing students for particular types of responses, and asking follow-up questions that push students to improve the clarity and quality of their scientific explanations. These types of feedback have been positively associated with student learning (Hattie and Timperley 2007; Kluger and DeNisi 1996), and are central to many definitions of quality formative assessment (Wiliam 2007).

These preceding formative assessment abilities are summarized in Table 1. Each of these abilities can differ in quality along a continuum from more traditional instruction to that aligned with effective formative assessment practice.

Table 1 highlights crucial differences between traditional, teacher-led instruction and formative assessment activities and instructional practices. These differences mean that many teachers with more traditional orientations toward instruction and student learning struggle to realize formative assessment in their classrooms (e.g. Furtak 2012; Atkin et al. 2005; Heritage et al. 2009). Indeed, research has indicated that the quality of teacher-designed formative assessment tasks varies considerably (Kang et al. 2014). Students’ relative and absolute understandings are subject to misinterpretation (Herppich et al. 2014); furthermore, teachers are likely to view student ideas as being correct or incorrect, rather than as a range of progressing ideas and conceptions (Furtak 2012; Coffey et al. 2011). Short- and long-term professional development programs intended to develop teachers’ formative assessment practices have yielded mixed results (e.g. Yin et al. 2008; Atkin et al. 2005; Falk 2012; Wiliam et al. 2004), and even teachers who have participated in such programs are not always equipped to provide effective feedback tailored to student thinking (Heritage et al. 2009).

Professional development in support of formative assessment abilities

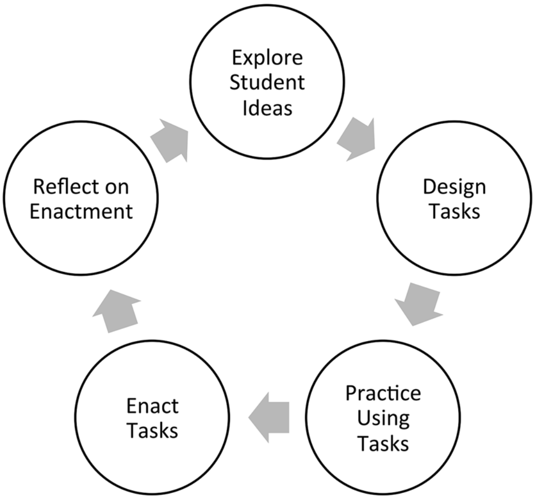
We have created a professional development model for the purpose of supporting teacher learning of the formative assessment abilities described in the preceding section. The aim is to support the transition from more traditional models of instruction toward formative assessment tasks and practices. Our approach builds on established approaches to assessment design that highlight the importance of identifying the construct to be assessed, collecting evidence of student performance relative to that construct, and then making interpretations of and inferences about what students know and are able to do on the basis of that evidence (Ruiz-Primo et al. 2001; NRC 2001a).

Researchers have hypothesized that teachers’ formative assessment abilities may be supported by representations of how student ideas develop in a domain (Bennett 2011; Heritage 2008). Learning progressions are one type of representation that is increasingly common in science education research. Learning progressions describe the pathways that students are likely to follow as they learn about disciplinary core ideas and practices (Corcoran et al. 2009), anchored on one side by “what is known about the concepts and reasoning of students entering school” (NRC 2007, p. 219) and at the other end by "societal expectations (values) about what society wants [middle] school students to understand about science" (p. 220). In the middle, learning progressions suggest intermediate understandings that are "reasonably coherent networks of ideas and practices and that contribute to building a more mature understanding." (p. 220).

The question remains, however, as to how teachers might use learning progressions in long-term professional development to support their understanding of student ideas, their formative assessment task design, and their abilities to draw out and respond to student thinking. Prior studies have established the importance of teachers engaging in long-term, discipline-specific professional learning experiences to support enduring changes in their classroom practices (Whitcomb 2013). As such, we created a professional development approach that incorporated elements of established models of effective, long-term professional development, including cycles of *planning*, *teaching* and *reflecting* (Borko et al. 2008), reflecting on evidence of teaching together (Ball and Cohen 1999; Sherin 2004), engaging in active learning strategies as well as explicit instruction to learn new instructional approaches (Penuel et al. 2011), and guiding teacher learning through active facilitation (Gröschner et al. 2014).

Our model, which we call the Formative Assessment Design Cycle (FADC; Furtak and Heredia 2014), is a five-step approach for professional development to support teachers in the development of formative assessment tasks with the support of a learning progression (Fig. 1). The cycle begins with a facilitator guiding teachers to *Explore Student Ideas* as well as their own understandings about the scientific concept to be taught (Borko 2004). In the second step, teachers *Design Tasks* collaboratively with their colleagues to elicit more and better information about student ideas during instruction. In the third step, teachers *Practice Using the Tasks* by rehearsing how they will enact formative assessment tasks together. The fourth step has the teachers *Enact the Tasks* during their instructional units and collect student work. Their enactment is videotaped. Finally, teachers *Reflect on Enactment* by exploring examples of student work, watching videotaped enactment of the formative assessment, and reflecting on what students learned (Sherin and Han 2002), as well as how to improve the formative assessment tasks and their accompanying classroom practices in the future.

Fig. 1 Formative assessment design cycle



The FADC as described above is an intervention intended to develop teachers' formative assessment abilities as defined in Table 1 in the following ways: We expected teachers to become more proficient at interpreting ideas along a continuum, and representing student thinking in ways consistent with the learning progression that underlay the study through exploring student ideas with the support of a learning progression. We anticipated that teachers would become more proficient at designing quality formative assessment tasks that would draw out student ideas by collaborating with the facilitator and colleagues in the process of formative assessment design. Finally, by practicing using these tasks with their colleagues, we expected that teachers would rehearse asking the types of questions that would elicit student thinking, and be better prepared to respond to these ideas with quality feedback during instruction.

Research questions

We set out to empirically test the relationship between teachers' participation in the FADC, their formative assessment abilities, and student achievement by conducting a multiple-year intervention study in which biology teachers from three high schools engaged in monthly meetings following the FADC, for three academic years. We collected measures of their formative assessment abilities, and assessed the achievement of two cohorts of their students: those in the baseline year of the study, and those from the third year of the intervention. Specifically, we responded to the following research questions:

1. To what extent does participation in the FADC support increases in the quality of teachers' formative assessment task design, questions to elicit student thinking, interpretation of student ideas, and feedback?
2. To what extent does teachers' proficiency in these formative assessment abilities predict changes in student achievement?

Method

Participants and setting

We partnered with three schools located in the same district outside a large city in the western US. The schools differ substantially in terms of student population and achievement. School 2 had a student population that was nearly 80 % Latino or Hispanic, with the same percentage of students receiving free or reduced lunch; students had test scores lower than state averages. In contrast, School 3 had a student population that is 75 % White students, fewer students receiving free and reduced lunch, and higher student achievement. School 1 was between Schools 2 and 3 on these measures.

We recruited all teachers who taught at least one 10th-grade biology class at each of these three schools, a total of 12 teachers during the baseline year. We assigned each teacher a numeric code, with the first digit indicating their school (1, 2, or 3) and the second digit indicating the individual teacher. After the baseline year, two of these teachers (Teacher 11 and Teacher 13) took jobs at different schools and a third (Teacher 21) left the profession, leaving nine teachers who completed all years of the study. These teachers ranged in experience from 4 to 21 years ($M = 12$, $Median = 10$), and the majority ($n = 7$)

had undergraduate degrees in Biology, with the remainder holding undergraduate major or minor degrees in physical science. All but two held Master's level degrees in Education. Four of the participants were males. All of the students[1] in the teachers' biology courses participated in the study (Baseline $N = 417$, Year 3 $N = 472$).

Design and procedure

Each year of the study had a one-group pretest–posttest design (Campbell and Stanley 1966), shown in Fig. 2. The nine teachers participated in baseline data collection at the beginning of the study, and their students that year took pre-posttests at the beginning and end of that baseline year. Then, beginning in the next academic year, teachers participated in monthly, on-site professional development meetings for three academic years. In the final year of the study, they again participated in measures of formative assessment ability, and their new cohort of students took the pre-posttest on the assessment of natural selection.

The intervention featured teachers participating in monthly, on-site meetings aligned with the FADC (Fig. 1). These meetings were conducted at each school site for about 60–90 min during teachers' common planning time, which happened before school at two sites, and during the school day at a third site. In total, teachers at each school participated in about 30 meetings over the course of three years. At each step of the FADC, teachers relied upon the Elevate learning progression (Furtak et al. 2014) to guide them in learning more about student thinking in the domain of natural selection (Fig. 3).

Each year of the study, the first meeting at each school began with teachers examining reports of student performance relative to the learning progression, and teachers identifying areas of their curriculum to focus upon for their formative assessment design (*Explore Student Ideas*). At subsequent meetings, teachers read articles about student thinking in this domain, examined their existing curriculum materials, and then designed formative assessment activities for their students (*Design Tasks*). They practiced using the activities with each other, envisioning how they would facilitate their activities with students, and anticipating the types of feedback they would provide if different types of student ideas were surfaced in class (*Practice Using the Tasks*). Then, once teachers had enacted the activity in their classrooms and been videotaped doing so (*Enact the Tasks*), the facilitator guided them to look at student responses together, as well as videotapes of classroom enactment (Borko et al. 2008), to consider the format of the activity, how it might be improved, and how they might improve both the activity design and their facilitation of the activity in subsequent years of the study (*Reflect on Enactment*).

Every meeting involved teachers working closely with a learning progression that represented student thinking in the conceptual domain of natural selection. The Elevate learning progression (Furtak et al. 2014) represents the multiple facets of a well-articulated explanation for natural selection, beginning with biotic potential, moving through the genetic origins of variations within populations of organisms, differential survival and reproduction of individuals on the basis of these variations, and changes in the distribution of these variations over time. Each column of the learning progression identifies one element of a well-articulated explanation of natural selection, with the top level

[1] Our human subjects agreement with the partner school district prohibited us from collecting student-level demographic data.

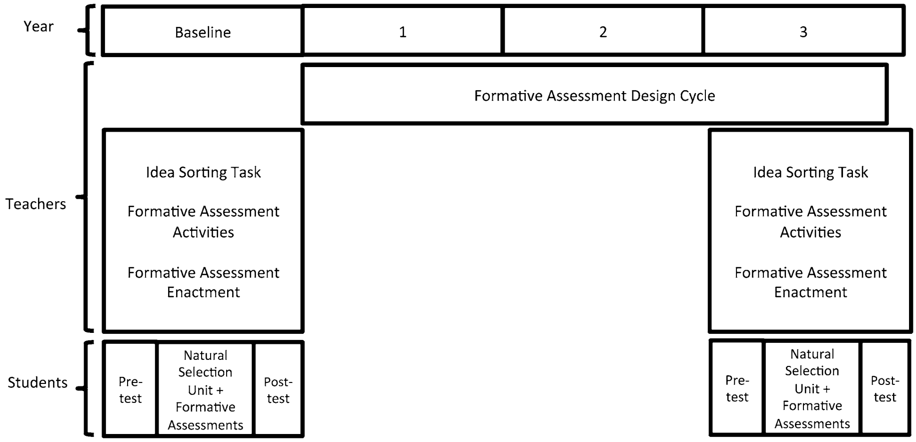
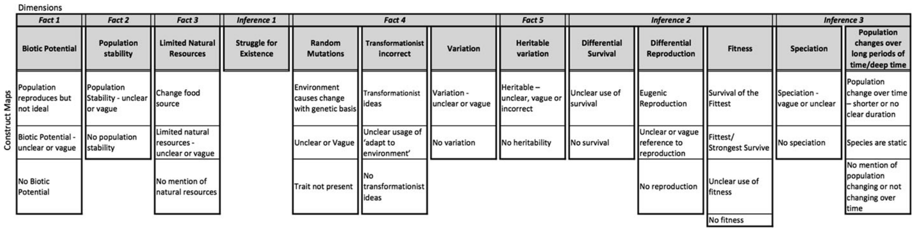

Year
Baseline
1
2
3
Formative Assessment Design Cycle
Teachers
Idea Sorting Task
Formative Assessment Activities
Formative Assessment Enactment
Idea Sorting Task
Formative Assessment Activities
Formative Assessment Enactment
Students
Pre-test
Natural Selection Unit + Formative Assessments
Post-test
Pre-test
Natural Selection Unit + Formative Assessments
Post-test

Fig. 2 Study design



Dimensions	Fact 1	Fact 2	Fact 3	Inference 1	Fact 4			Fact 5	Inference 2			Inference 3	
	Biotic Potential	**Population stability**	**Limited Natural Resources**	**Struggle for Existence**	**Random Mutations**	**Transformationist incorrect**	**Variation**	**Heritable variation**	**Differential Survival**	**Differential Reproduction**	**Fitness**	**Speciation**	**Population changes over long periods of time/deep time**
Construct Maps	Population reproduces but not ideal	Population Stability - unclear or vague	Change food source		Environment causes change with genetic basis	Transformationist ideas	Variation - unclear or vague	Heritable – unclear, vague or incorrect	Unclear use of survival	Eugenic Reproduction	Survival of the Fittest	Speciation - vague or unclear	Population change over time – shorter or no clear duration
	Biotic Potential - unclear or vague	No population stability	Limited natural resources - unclear or vague		Unclear or Vague	Unclear usage of 'adapt to environment'	No variation	No heritability	No survival	Unclear or vague reference to reproduction	Fittest/ Strongest Survive	No speciation	Species are static
	No Biotic Potential		No mention of natural resources		Trait not present	No transformationist ideas				No reproduction	Unclear use of fitness		No mention of population changing or not changing over time
											No fitness		

Fig. 3 Elevate learning progression

representing the 'correct' response; lower levels articulate common misconceptions about each dimension (Fig. 3).

Teachers used the learning progression to help them set goals for designing their formative assessment task, and then used it again when interpreting student responses to their assessments when reflecting upon student work. Since each school met independently to engage in the FADC, these sets of formative assessment activities differed across schools, but were common within schools.

Measures and sources of data

We operationalized our conceptualization of formative assessment into measures aligned with the four abilities described above: designing formative assessment tasks, asking questions to elicit student ideas, interpreting student ideas, and providing feedback which moves student thinking forward. We collected multiple sources of data in the Baseline and Year 3. We describe these sources of data alongside their corresponding measures below.

Formative assessment task ratings

We measured the extent to which teachers were able to *design formative assessment tasks* with a six-item rating system based on prior research on assessment task design (e.g. Kang et al. 2014; Ruiz-Primo et al. 2001) that evaluated the activities on a scale of 0 (traditional) to 5 (consistent with quality formative assessment). Items rated the *outcome space of the activity*, the *type of instruction* that might accompany the activity, the *type of knowledge* the activity elicited, the *type of information* about student ideas the assessment was designed to provide, the *potential of the activity to make students' scientific understandings visible*, and the *ease of interpreting* these understandings (See Appendix 1 for full listing of items). Experienced biology teachers ($N = 6$) who had previously worked with the authors of this study around formative assessment, but who did not participate in the study, rated the activities on each of these six items; intraclass correlations (ICC) for the teachers' ratings ranged from 0.80 to 0.96. We generated a variable for the quality of each teacher's formative assessment tasks by calculating his or her mean task rating for the Baseline year, and the mean task rating for Year 3 (theoretical min = 0, max = 5).

Interpretation of student ideas

Each teacher completed a sorting task (cf. Friedrichsen and Dana 2003; Smith et al. 2013) at the beginning and end of the study. The sorting task asked teachers to read an assessment activity about *Biston betularia* moths in England during the industrial revolution. Teachers

Table 2 Coding System for teacher eliciting questions and feedback

Score	Code	Description	Example
Teacher eliciting questions			
0	Not present	Segment does not contain an eliciting or surfacing question or statement.	
1	Instructional	Cued elicitation of students' ideas; teacher asks questions while providing heavy clues for the information being sought.	"What's the answer to number 2?" "What are we dealing here with?… natural…"
2	Eliciting student thinking	Teacher asks students to share their ideas, conceptions, opinions or interpretations.	"What do we mean by evolution?" "What are your ideas why the moth population changed in that way?" "Who has an idea about the process behind it?"
Feedback			
0	Not present	Teacher response contains no response to student idea.	
1	Evaluative or limiting	Teacher response is evaluative, or teacher interrupts student and finishes student thought.	"Yes." "No." "Right." "Good job."
2	Neutral	Teacher repeats or revoices students' words, other neutral response.	"Okay." "Yeah." "I see." "So you're saying that…"
3	Pushing student thinking	Promotes students' thinking by asking them to elaborate their responses or by asking for more information about the previous question; Provides descriptive or helpful feedback about quality of student idea.	Following up with a why/how question, disagreeing with student's response, asking "what does that mean?" "The fact that you said, like, 'pass on those genes' made that a really good answer."

were then provided with seven actual student responses to this assessment, and asked to sort them in a way that made sense to them (Appendix 2).

We posed the same task to a research team member who was trained in scoring students' ideas relative to the Elevate learning progression as part of previous studies (e.g. Furtak 2012); nomination of this team member was in accord with criteria suggested in Palmer et al. 2005 (4 years of teaching experience, knowledgeable researcher in the domain of biology education and nominated by the research team); and recorded her sorting of ideas. Each teacher's idea sorting score was then established as the direct agreement between their categorization of student ideas and the categorization of ideas done by the researcher relative to the learning progression. Finally, we transformed each teacher's score to a scale from zero (no agreement with sorting according to learning progression) to 1 (exact agreement).

Teacher eliciting question and feedback coding system

To measure the quality of the *questions teachers asked to elicit student ideas*, and the *quality of verbal feedback teachers provided to students*, we applied two coding systems to teachers' questions and feedback to student ideas adapted from previous analyses of the quality of teacher talk moves in formative assessment classroom discussions (e.g. Ruiz-Primo and Furtak 2006, 2007; Seidel and Prenzel 2006), shown in Table 2.

We videotaped each teacher on multiple occasions enacting the formative assessment tasks collected as artifacts in the Baseline and Year 3 of the study. These videotapes were made with a single camera positioned at the side or back of the classroom with a boom or lapel microphone used to capture the teacher and students' voices. To track alignment between talk formats more consistent with quality formative assessment as compared to traditional instruction, we identified all instances in the videotapes of whole-class discussions, and then segmented each of those discussions by talk turn. We then performed in-depth analyses of these discussions in the Videograph program (Rimmele 2015) using the coding system described in Table 2. Two raters independently coded 20 % of the 89 total videos and established acceptable levels of agreement with Cohen's κ as follows: teacher question = 0.90; teacher feedback = 0.85. The remaining videos were divided among raters and coded independently.

We assigned a numeric value to each code reflecting its quality, with increasing values representing higher quality questions and feedback, as indicated in Table 2. This means that each instance of a teacher asking a question was treated as an occasion that could be assigned a particular score (min = 1, max = 2), and each instance of a teacher providing feedback could be treated as an occasion that could be assigned a particular score (max = 1, max = 3). Then we created averages across occasions to generate variables for mean question quality and mean feedback quality in the Baseline and Year 3.

Daphne assessment of natural selection

We assessed student achievement with the Daphne Assessment of Natural Selection (DANS), which consists of 17 ordered multiple-choice items (Briggs et al. 2006) aligned with the Elevate Learning Progression, each of which frames natural selection in a variety of plant and animal contexts. Since the same test was used pre and post through the course of the study, we kept items secret from the teachers until the conclusion of the study. We scored the items dichotomously (theoretical min = 0, max = 17) and then calculated internal consistency at each administration with Cronbach's alpha (Baseline pretest

$\alpha = 0.34$, posttest $\alpha = 0.61$; Year 3 pretest $\alpha = 0.35$, posttest $\alpha = 0.62$). Tests comprised of ordered multiple-choice items can have alphas in this lower range (Alonzo and Steedle 2009) because of their multidimensionality; furthermore, the fact that the alphas change so much from pre to posttest indicates a homogenous sample at the pretest because students have not yet been exposed to the curriculum the test is designed to assess, but much of this homogeneity disappears at the posttest as students experience differential learning gains. To be conservative, we proceed by only interpreting group averages, rather than individual student scores. More information about the DANS, its construction, and alignment with the content of the study can be found in Furtak et al. (2014).

Analytic model

Our data suggests an analysis that will allow us to examine the nested nature of the data (i.e., students within teachers) as well as the relationship of our four variables of formative assessment abilities with student achievement. As such, we modeled the relationship among teachers' formative assessment abilities and student achievement through two Hierarchical Linear Models (HLM) that estimated the contribution of each of the variables of formative assessment abilities to students' posttest scores in the Baseline and Year 3. The HLM models examined the extent to which variance in student achievement before and after the teachers' natural selection units could be attributed to individual differences at the student level (Level 1) as well as differences in the context of those students' learning at the teacher level (Level 2; Bryk and Raudenbush 1992); the two separate HLM analyses allowed us to analyze the Baseline and Year 3 data separately. We conducted multilevel models using the lme4 package (Bates et al. 2012) for R (R Core Team 2012). Posttest scores on the DANS were the outcome variables for the study, with pretest scores serving as a student-level predictor (as commonly observed, we expected that the pretest results would be positively associated with posttest results). Teacher-level predictors included variables for quality of formative assessment task design, asking questions to elicit student thinking, interpretation of student ideas, and quality of feedback to student ideas. We expected that these teacher level variables would predict mean posttest scores positively, even after pretest scores are controlled in the student level model. The Level 2 sample size was small (only 9 teachers) and, as such, the standard errors of the second-level variances are underestimated; however, Maas and Hox (2005) found that there is no support for a bias in the regression estimates (see also Schoppek 2015).

We specified the following Level 1 model predicting that the achievement of a student i in teacher j (Y_{ij}) is a function of the teacher intercept (b_{0j}) plus a component that reflects the linear effect of student pretest score (b_{1j}) plus random error (e_{ij}).

$$Y_{ij} = b_{0j} + b_{1j}\left(\text{pretest}_{ij}\right) + e_{ij} \tag{1}$$

Our Level 2 model then posited that each group (teacher)'s intercept (b_{0j}) is a function of a common, fixed intercept (β_{00}) plus the linear effect of each of the teacher-level variables plus a random between-group error (u_{0j}). The slope of pretest across groups is specified to be fixed.

$$\begin{aligned} b_{0j} = {} & \beta_{00} + \beta_{01}(\text{quality of formative assessment task design}) + \beta_{02}(\text{eliciting questions}) \\ & + \beta_{03}(\text{interpreting ideas}) + \beta_{04}(\text{teacher feedback}) + u_0 \end{aligned} \tag{2}$$

$$b_{1j} = \beta_{10} \tag{3}$$

The combined multilevel model with one student level and four teacher level explanatory variables is shown in Eq. (4):

$$\begin{aligned} Y_{ij} = \beta_{00} + \beta_{10}(\text{pretest}) + \beta_{01}(\text{quality of formative assessment task design}) \\ + \beta_{02}(\text{eliciting questions}) + \beta_{03}(\text{interpreting ideas}) \\ + \beta_{04}(\text{teacher feedback}) + u_{0j} + e_{ij} \end{aligned} \tag{4}$$

Results

We present our results by research question. We begin by presenting results of our analysis for each the four measures of formative assessment abilities (interpretation of student ideas, quality of formative assessment task design, quality of questions eliciting student thinking, and quality of responses to student ideas) separately, and then explore the relationship among these measures within and across teachers. Finally, we relate these measures to student learning.

(1) What is the relationship between teachers' participation in the FADC and the quality of their formative assessment task design, questions to elicit student thinking, interpretation of student ideas, and feedback?

We summarize each teacher's formative assessment abilities in Table 3, and discuss each below.

Table 3 Summary of measures of formative assessment ability, by year

School	Teacher	Task design (min = 0, max = 5)		Eliciting questions (min = 0, max = 2)		Idea interpretation (min = 1, max = 2)		Feedback (min = 1, max = 3)	
		Baseline	Year 3	Baseline	Year 3	Baseline	Year 3	Baseline	Year 3
1	12	2.67	4.12	1.25	1.33	0.57	0.57	1.95	2.28
	14	3.37	4.12	1.28	1.67	0.43	0.29	2.17	2.11
	15	3.91	3.83	1.19	1.29	0.43	0.57	1.58	1.88
2	22	2.39	3.73	1.15	1.06	0.29	0.57	1.89	2.25
	23	4.22	3.71	1.38	1.36	0.17	0.57	1.83	2.34
3	31	4.58	3.70	1.08	1.38	0.71	0.86	2.11	2.24
	32	4.58	3.70	1.08	1.45	0.29	0.86	2.12	2.19
	33	1.22	3.70	1.08	1.35	0.57	0.71	2.13	2.24
	34	1.19	3.70	1.46	1.52	0.57	0.71	2.52	2.31
Mean		3.13	3.81	1.22	1.39	0.45	0.63	2.03	2.20
SD		1.33	0.18	0.14	1.67	0.17	0.18	0.26	0.14

Quality of formative assessment task design

Teachers used collectively developed activities at schools 1 and 2, and accompanied those common activities with their own materials, which led to variations in quality of formative assessment task design among teachers. In contrast, teachers at school 3 converged with a task quality of 3.40, since all four teachers used the same activities. However, at school 3, the two teachers with lower-quality formative assessment tasks in the baseline year appeared to benefit from co-designing formative assessment activities with their colleagues who had higher-quality baseline tasks (e.g. Teachers 33 and 34). At the same time, teachers who had higher-quality baseline tasks actually had their task quality decrease in Year 3 (e.g. Teacher 33 and 34). Table 3 presents a positive picture of teachers' progress in the design of high-quality formative assessment tasks. The overall mean task quality increased from the baseline to Year 3, and the standard deviations drastically decreased. However, this difference was not significant ($t(8) = -1.53$, $p = 0.16$).

Asking questions to elicit student ideas

The mean quality of questions asked during the videotaped lessons increased for the majority of teachers (Table 3). The communities of practice at school 1 and school 3 seem to have been especially successful in this regard, whereas the two teachers at school 2 decreased slightly in the quality of questions asked. The mean quality of questions in Year 3 was statistically significantly higher than in the baseline year ($t(8) = -2.79$, $p = 0.02$).

Interpretation of Student Ideas

Teachers varied in their interpretation of student ideas as compared to the learning progression (Table 3). All teachers' interpretations of student ideas were in greater agreement with the learning progression in Year 3 with the exception of two teachers, one who stayed the same (Teacher 12), and another who decreased (Teacher 14). The mean Year 3 interpretation of student ideas was statistically significantly higher than in the Baseline year ($t(8) = -2.68$, $p = 0.03$), suggesting that the majority of the teachers came to sequence students' ideas as a continuum aligned with the learning progression.

Quality of teachers' responses to student ideas

All but teachers 14 and 34 increased the quality of the feedback they provided to students, and they decreased only slightly (Table 3). Overall, results indicate decreasing variance in the quality of feedback provided to students accompanying the overall increase in the quality of feedback over the course of the study. This increase was statistically significant ($t(8) = -2.28$, $p = 0.05$) and suggests that teachers' practices converged through their collaborative participation in the FADC.

Profiles of changes in formative assessment abilities

Before going into a more detailed descriptive analysis of the profiles of change for the different teachers at the different school sites, we again point out that positive developments in teachers' formative assessment abilities occurred across all teachers and all sites, and patterns emerged within schools. This result is likely attributable to the intervention, as

Table 4 Student achievement by school and teacher for the baseline and year 3 of the study

Teacher	Year	N	Pretest				Posttest				Mean prepost gain	Mean percent gain	Effect size
			Min	Max	Mean	SD	Min	Max	Mean	SD			
12	1	31	2	11	6.8	2.2	1.0	13.0	7.2	2.7	0.4	0.02	0.16
	4	51	3	13	7.2	2.2	5.0	16.0	9.4	5.2	2.2	0.13	0.55
14	1	62	1	13	6.4	2.2	3.0	14.0	7.8	2.7	1.5	0.09	0.57
	4	76	1	12	6.1	2.0	0.0	14.0	7.2	2.5	1.2	0.07	0.49
15	1	29	2	11	6.5	2.1	1.0	11.0	6.0	2.6	−0.4	−0.02	−0.21
	4	23	3	10	5.4	1.9	2.0	11.0	7.0	2.8	1.5	0.09	0.67
22	1	23	3	11	5.4	2.1	1.0	9.0	5.8	2.2	0.4	0.02	0.19
	4	23	0	10	5.9	2.5	3.0	11.0	6.1	2.2	0.2	0.01	0.08
23	1	14	0	10	5.3	2.6	2.0	9.0	6.1	2.3	0.8	0.05	0.33
	4	50	1	10	5.7	2.0	2.0	11.0	6.7	2.1	1.0	0.06	0.49
31	1	109	1	14	7.5	2.4	3.0	17.0	8.6	3.0	1.2	0.07	0.40
	4	93	1	12	6.5	2.2	2.0	16.0	8.1	2.9	1.6	0.09	0.62
32	1	100	2	13	7.2	2.2	0.0	14.0	8.0	2.7	0.8	0.05	0.32
	4	76	1	14	6.9	2.1	3.0	16.0	8.8	3.1	1.8	0.11	0.72
33	1	67	2	14	7.6	2.7	3.0	15.0	8.4	2.7	0.8	0.05	0.30
	4	20	3	10	6.0	1.9	5.0	13.0	8.1	2.4	2.1	0.12	0.97
34	1	37	4	11	7.4	1.9	1.0	16.0	8.6	3.3	1.2	0.07	0.45
	4	16	3	10	6.4	1.9	4.0	13.0	7.8	2.9	1.3	0.08	0.57

teachers collaboratively designed and rehearsed formative assessment tasks, learned to interpret student responses relative to the same representation of student ideas about natural selection, and anticipated student responses and likely feedback. At the same time, within each school, we observed patterns of change specific to individual teachers.

At School 1, all teachers increased in the quality of task design and eliciting questions; Teacher 12 stayed the same in idea interpretation, and increased in Feedback quality while Teacher 15 increased on both of these measures. However, Teacher 14 decreased in both idea interpretation and feedback quality. This result suggests that teacher collaboration at School 1 supported quality task design as well as questioning practices, with variations in idea interpretation and feedback quality.

At School 2, we observed a less promising pattern of change. While both teachers 22 and 23 increased in their idea interpretation and feedback quality, and Teacher 2 increase in the quality of task design, we observed decreases in eliciting questions for both teachers, and a decrease in task quality of Teacher 23. Teachers at School 2 did not use the same activities in the final year of the study, suggesting that the school-based collaboration did not necessarily support uniform changes in task quality for both participating teachers.

Finally, at School 3, we observed increases in eliciting questions and idea interpretation for all teachers, and an increase in feedback quality for all but Teacher 34. Interestingly, and as noted above, we saw two teachers enter the study with lower-quality tasks (Teachers 33 and 34) and two with higher-quality tasks (Teachers 31 and 32), and the teachers 'met in the middle' in Year 3, leading to a decrease in task quality for Teachers 31 and 32, but an increase for Teachers 33 and 34. This result suggests that the department of biology teachers overall benefitted from the study in terms of task design, but this came at the expense of the higher-quality tasks of two members of the department.

These many patterns of change, particular to teachers within specific schools, were in some cases large, some small, some negative, and some positive; as such, they raise the question as to how these variations in abilities were predictive of changes in student achievement in the Baseline and Year 3 of the study.

(2) How are teachers' formative assessment abilities related to student achievement?

We now turn our analysis to determining the relations of these variations in teachers' formative assessment abilities to student achievement. We remind the reader that although our study followed the same teachers for multiple years, the students these teachers were instructing were different in the Baseline and Year 3 of the study. We controlled for this difference by using and interpreting pretest scores as measures of prior knowledge.

Descriptive statistics presenting the mean pre and posttest scores in the Baseline and Year 3 of the study are provided in Table 4; effect sizes indicate greater gains in pre-post achievement between the students in the Baseline and Year 3 of the study for all but Teacher 14 and 22.

As a first step, we determined if students in Year 3 learned more as compared to students in the baseline year of the study. To make this determination, we ran a student-level analysis on the posttest scores with the pretest as a covariate to test for significant differences within different years. The ANCOVA analyses using the whole dataset showed a significant effect of year when controlling for the pretest, as well as dummy codes for teacher and school, on the students' natural selection achievement on the posttest ($F_{(6.890)} = 2.45$, $p < 0.05$). This result suggests that students in Year 3 had statistically significantly higher achievement gains than students in the Baseline year of the study.

Next, we turned to a multilevel analysis of the data. Since our study measured the achievement of two different cohorts of students—those in a baseline year, and those after teachers participated in 3 years of the professional development intervention—we modeled student achievement with two separate HLM models. We standardized all variables ($M = 0$, $SD = 1$) before entering them into the multilevel analysis. First, we checked the proportion of total variance in the outcome (i.e., posttest scores) that can be explained by group membership (i.e., teachers), i.e., the intraclass correlation (ICC). As suggested by Lee (2000), multilevel models are useful when the ICC is more than trivial (i.e., greater than 0.10). Results of the unconditional model showed that the between-group ICC for the posttest in the baseline year was 0.12 and 0.15 in Year 3; put differently, there is considerably more variance in posttest scores within teachers then there is between them, but still a nontrivial amount variation of between teachers. Therefore, we proceeded with a multilevel model.

We selected the random intercept model for the data. We also examined the significance of the random slope of the pretest as described in Snijders and Bosker (2012) and found that, while there was some slope variation across teachers, the joint test of variance and covariance based on the likelihood ratio test[2] was non-significant for both baseline and Year 3 ($p_{\text{baseline}} = 0.26$ and $p_{\text{Year3}} = 0.20$). This finding is also in agreement with Schoppek (2015) who could not establish significant benefit of a random slope model in cases with small sample sizes.

The results of our HLM analysis are summarized in Table 5. The model indicates that the mean overall posttest value (fixed intercept) was zero. As expected due to standardized variables, pretest scores significantly positively predicted posttest scores in both years ($\beta_{10} = 0.33$ and $\beta_{10} = 0.40$ respectively). These standardized coefficients refer to the expected difference in posttest scores associated with a one-standard deviation difference in pretest scores.

After controlling for this individual-level relationship in the baseline year model, we did not find a significant association between task quality, eliciting questions, or idea interpretation and the posttest scores. This result is not surprising given that the nine teachers had not yet learned about the learning progression or formative assessment designs that were the centerpiece of the study, and so variations in the quality of these formative assessment abilities were not yet predictive of student achievement. However, there was a significant association between feedback quality and the posttest scores ($\beta_{04} = 0.21$), indicating that the quality of responses teachers gave to student ideas was high enough to be the only significant contributor to student learning in the baseline year among the formative assessment ability variables.

After three years of participation in the FADC, the pattern of associations in the HLM analysis changed. The significant and positive association between feedback and the posttest scores disappeared despite significant differences in mean feedback quality from Baseline to Year 3 of the study. As Table 3 indicates, the variance in feedback quality decreased from the Baseline to Year 3, reflecting an increase in mean feedback quality that was consistent across most teachers. Thus the increased homogeneity in feedback quality at in Year 3 indicates a positive outcome of the professional development, but did not explain variance in student outcomes at the posttest in the HLM model.

In contrast, the quality of tasks teachers designed ($\beta_{01} = 0.38$) and teachers' interpretation of ideas ($\beta_{03} = 0.42$) had positive and significant contributions to posttest scores in Year 3. The variance in the quality of teachers' tasks decreased drastically from the Baseline to Year 3 of the study, indicating overall changes in task quality; however, task quality did not uniformly increase, a heterogeneity that was associated with differences in student posttest scores in Year 3. We observed significant differences in mean idea

[2] Note that the likelihood ratio tests may be preferable when there are fewer groups.

Table 5 Results of HLM analysis

	Model 1 baseline coefficient (β)	Model 2 Year 3 SE	Coefficient (β)	SE
Fixed effect estimates				
Intercept	0.00	0.04	0.00	0.05
Pretest scores	0.33*	0.04	0.40*	0.04
Task quality	0.05	0.05	0.38*	0.08
Eliciting questions	−0.07	0.05	0.01	0.05
Idea interpretation	0.05	0.04	0.42*	0.08
Feedback	0.21*	0.05	−0.03	0.05
Random effect estimates				
Residual standard deviation L1	0.91		0.93	
Standard deviation L2	0.008		0.056	
R^2 (Level 1, Level2)[a]	10.2, 99.9		16.5, 97.9	

$N_{baseline}N_{baseline} = 472$, $N_{Y3} = 428$ * $p < 0.05$

[a] The R^2 value represents how much variance is explained as the proportion of variance modeled by explanatory variables. In the current analysis, there is unexplained variance at two levels, hence we calculated R^2 for two levels

interpretation, and this variance became predictive of posttest scores in the Year 3 HLM model.

These findings suggest that the extent to which teachers were able to design quality formative assessment tasks and interpret ideas in alignment with the learning progression predicted student achievement positively in Year 3; at the same time, nearly all the teachers improved in their feedback practices, and so variance in their responses to student ideas was no longer significantly predictive of student posttest scores. Despite mean increases from the Baseline to Year 3 of the study, no significant relationship was found for the mean quality of teacher questions in either HLM model.

We interpret these findings as follows: Teachers' task quality varied considerably at the beginning of the study, but this variance wasn't as important as the quality of feedback teachers were giving to students. This result is consistent with studies that have suggested the overall importance of high-quality feedback in supporting student learning (Hattie and Timperley 2007). In contrast, at the conclusion of the study, the means of all formative assessment ability variables increased, with significant differences in mean question and feedback quality; however, variability in feedback was no longer significantly predictive of student posttest scores. This suggests that the increased quality of formative assessment tasks and teachers' ability to interpret student ideas became more important, once the quality of feedback practices increased for most teachers.

Discussion

In this paper, we explored changes in teachers' formative assessment abilities as captured by the *quality of their formative assessment task design*, *questions to elicit student thinking*, *interpretation of student ideas*, and *feedback*, and the influence of these abilities on student achievement. We measured these abilities prior to and after teachers had participated in long-term, school-based collaborative professional development centered on a learning progression for natural selection over the course of three academic years.

On average, we observed significant increases in teachers' question quality, interpretations of student ideas, and feedback quality, but not task quality. These results suggest the efficacy of the three-year professional development intervention in supporting increases in some, but not all, of the formative assessment abilities in the study. Results of our multilevel models indicate that, while feedback quality was significantly predictive of student posttest scores in the baseline year of the study, teachers' task quality and interpretation of student ideas contributed significantly positively to student achievement gains in Year 3. We interpret these findings in light of previous studies of formative assessment, professional development, and classroom discourse.

First, the positive and significant contribution of feedback to student posttest scores in the baseline year is consistent with prior research syntheses indicating the importance of feedback in supporting student learning (Black and Wiliam 1998; Hattie and Timperley 2007). Furthermore, the fact that we observed significant increases in teacher feedback across the course of the study suggests that teachers' participation in professional development can support the development of this teaching practice (Kiemer et al. 2015). The result that feedback quality did not relate significantly to student posttest scores in Year 3 may be interpreted as follows: the nearly uniform increase in this formative assessment ability across teachers meant that it no longer explained variations in student achievement.

Instead, student posttest scores were explained by different variables at the end of the study: task quality and idea interpretation. We observed significant increases in teachers' idea interpretation from the Baseline to Year 3, and whereas this ability was not significantly related to student posttest scores in the baseline year, it was in Year 3. One possible interpretation is that the use of a learning progression in the professional development meetings may have had an influence on the ways in which teachers interpreted student ideas during classroom enactment of formative assessment tasks. Although the design of the study was not able to isolate this aspect of the study, and does not allow us to make causal attributions, this finding supports the long-hypothesized link between learning progressions and teachers' interpretations of student thinking (e.g. Furtak 2009a; Bennett 2011; Heritage et al. 2009). It can be argued that through participation in the professional development intervention teachers shifted more towards a view of learning that acknowledges the importance of attending to students' everyday ideas (e.g. NRC 2001a). Furthermore, the significant contribution of teachers' idea interpretation to student achievement is supported by studies that have argued for the importance of teacher knowledge about student thinking—what Shulman called pedagogical content knowledge (Shulman 1986)—in supporting student learning (e.g. van Driel et al. 1998). Indeed, Falk (2012) argued that teachers' pedagogical content knowledge was very closely related with their formative assessment practice; that is, by conducting formative assessment, teachers built their pedagogical content knowledge. In turn, teachers drew upon that pedagogical content knowledge to further enact formative assessment. Of course, future studies will need to investigate this effect in 2-group experimental designs in order to systematically relate these findings to the use of a learning progression as a scaffold in professional development.

We also found that the quality of teacher-created formative assessment tasks also increased across the course of the study, and the variance in task quality explained a significant amount of student posttest scores in Year 3. This reflects the fact that some teachers gained more through the process of collaborative design, whereas others saw their task quality decrease, and these variations were more important in terms of student achievement as compared to the quality of teacher feedback in Year 3. We note that the variance in activity quality was somewhat restricted as the teachers at school 3 used exactly the same tasks across classrooms, while their colleagues at other sites used additional materials on top of their co-designed tasks, or slightly different versions of those activities. Future research may examine more closely the relationship between teachers' common formative assessment tasks and the quality of their classroom formative assessment practices. This result, in combination with our findings about teachers' classroom practices, also indicates that the teachers interactions with students were not directly related to the tasks they were using.

Finally, a key finding of our study was that the quality of teachers' questions, which did increase across the course of the study, did not make a positive contribution to student achievement in either the Baseline or Year 3. While it is almost universally acknowledged that open-ended questions are more likely to surface the nature of students' thinking (e.g. Coffey et al. 2011), questions alone may not be as instructive as targeted, informational feedback that has been shown to positively impact student achievement (Hattie and Timperley 2007). In fact, asking students open-ended questions alone may not support the development of their learning as much as asking students combinations of open and closed-ended questions and following them up with meaningful feedback (e.g. Dillon 1985; Smith and Higgins 2006). This alternating between what Scott, Mortimer, and Aguiar (2006) called the 'authoritative and dialogic functions' of discourse actually can mean that while asking open-ended questions may surface student thinking, following up with closed-ended questions that push students toward particular answers may be a more effective strategy to help them learn.

Future studies may more carefully explore the classroom climates of these classrooms in order to better understand the possible interaction between climate, teacher questions and student responses. Nonetheless, since teachers' use of questions has been said to be a highly routine practice that is very resistant to change (Oliveira 2010), the significant increase in question quality throughout the study suggests the efficacy of the intervention in supporting changes in teachers' classroom practices.

When interpreted in the context of other studies of teachers' learning in professional development, our results underscore that teachers have varying take-aways from these learning experiences. The transfer of new knowledge into teachers' classrooms is an individual process affected by various cognitive and motivational–affective aspects, as well as situational and organizational frameworks (Clarke and Hollingsworth 2002). Such individual trajectories of teacher learning have been identified in other studies (e.g. Thompson et al. 2013). Ultimately, making sense of student thinking and responding to those ideas with quality feedback showed the greatest contribution to student achievement (Black and Wiliam 1998; Hattie and Timperley 2007; Kingston and Nash 2011).

The findings of this study have important implications for the design and conduct of professional development, and the possible linkages between professional development, teacher formative assessment abilities, and student learning. Collaborative professional development following teaching cycles of *planning*, *teaching* and *reflecting* (Borko et al. 2008) and incorporating effective components of professional development (Desimone 2009; van Veen et al. 2012; Wilson 2013) can raise teacher effectiveness in a variety of schools with varying socio-cultural, social, and economic backgrounds. Furthermore, our study suggests that the scaffold of a learning progression in a particular conceptual domain can act as a scaffold for teacher interpretation of student ideas. These representations are increasingly being developed for different scientific domains and practices (Duschl et al. 2011) and are the foundation of the *Next Generation Science Standards* in the United

States (NGSS Lead States 2013). This widespread availability should be accompanied by sustained opportunities for teachers to learn about the ideas that they contain. Finally, our study suggests that engaging teachers in collaborative design of formative assessment tasks can raise the quality of tasks used across teachers and within schools, and can contribute positively to student learning.

We acknowledge multiple limitations of this study. First, the design of the study did not allow causal inferences to be made; had a control group been used, and with a sufficient sample size, we may have been able to identify further impacts of the FADC on teacher practice and student achievement. Furthermore, future research may further disaggregate the interaction between different groups of students and teachers' formative assessment abilities, which was not possible since we were not able to collect other possible student-level covariates such as age, gender, or socioeconomic status in this study. We also intend, in future work, to use more in-depth analyses of teachers' classroom data to better understand the ways in which those teachers used the formative assessment tasks they designed, as well as the quality of student ideas shared in whole-class conversations. Finally, the finding that teachers' interpretation of student ideas and feedback significantly contributed to variance in student learning suggests a connection of the specifics of student ideas, relative to the learning progression, that teachers responded to with feedback to increase students' learning. Future studies may more closely explore the disciplinary substance of student ideas in this dataset (Coffey et al. 2011).

In closing, we reflect that we have completed a complex study which, as a vestige of its design, was not intended to isolate and experimentally confirm the effect of formative assessment on student achievement. Rather, our study allowed us to track the development of teachers' formative assessment abilities across the course of multiple years, and then test the influence of those abilities on two separate cohorts of students. Overall, we are encouraged that teachers came to categorize student ideas in alignment with the learning progression; moreover, that the changes in teachers' idea interpretation and responses to student thinking contributed significantly to student posttest scores.

Acknowledgments The authors thank the special issue editors for their thoughtful feedback that greatly improved this manuscript, as well as the anonymous reviewers for their comments and suggestions. We are also grateful to Andy Maul and Derek Briggs for their technical advice with HLM analyses. This material is based upon work supported by the National Science Foundation under Grant No. 0953375. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix 1: Formative assessment task rating items

1. What type of space is there in the activity for students to share their ideas?

- 1—None/no space for student-provided answers (e.g. lecture slides)
- 2—Multiple-choice questions only
- 3—Fill-in-the-blank
- 4—Short-answer (can include multiple-choice plus justification)
- 5—Open, lengthy student-constructed responses (e.g. essay or student-constructed representation)

2. What is your impression of the type of instruction that would accompany this activity? Be sure to answer based on how the activity is written, not how you might use it.

   - 1—None/not applicable
   - 2—Lecture style of instruction
   - 3—Teacher-centered questioning
   - 4—Teacher-student dialogic interaction
   - 5—Student-led instruction or student–student interaction

3. What type of knowledge is being elicited/promoted/targeted?

   - 1—None/not applicable
   - 2—Declarative/factual knowledge/recall
   - 3—Procedures
   - 4—Schematic knowledge/the 'big ideas' of science
   - 5—Strategic/deciding when or how to use knowledge

4. What type of information about student ideas does this activity seem designed to provide?

   - 1—None/not applicable
   - 2—Correctness in a binary sense (e.g. right/wrong)
   - 3—Mostly correctness, but with the potential for a little information about student ideas
   - 4—Information about what students are thinking
   - 5—Complex information about student thinking, including common student everyday ideas/misconceptions

5. What is the potential of this activity to make students' scientific understandings visible?

   - 1—Not applicable
   - 2—None
   - 3—Low
   - 4—Moderate
   - 5—High

6. What is your impression of how difficult or easy it might be to interpret these scientific understandings?

   - 1—Not applicable/no clear request for information about student ideas in the activity
   - 2—Not sure
   - 3—Hard and slow to interpret
   - 4—Easy to interpret, but it would take some time to go through the answers
   - 5—Easy and fast to interpret

Appendix 2: Student idea sorting task

Interviewer questions for teacher:

1. I'd like you to take a minute to read through these student responses, and then sort them according to the ideas that they contain. Please talk aloud as you read and make your decisions so we know what you're thinking.
2. How would you describe the kinds of ideas represented in the responses?
3. What would you do in class to address the different ideas in the responses?
4. Some of these responses have multiple ideas in them; what does that indicate to you about what these students are thinking?

Sorting task—student responses

Number	Student idea
1	I chose a new answer because the color of the bark is what caused the genes in the moth to change from light to dark
2	The industrial pollution colored the tree back and the moths turned darker also to avoid being preyed on by birds. The moths had to adapt to their changing environments by changing color
3	I chose answer () because that is natural selection. The article said that the lighter moths were the main type of moth, meaning that there were also darker moths. Since the bark slowly became darker, it would take quite a long time for the dominant moth color to change. The darker moths would be producing more, and therefore they would slowly overpopulate the light moths
4	I chose () because the colors of the bark of trees was the selective force. Though the colors mutation of peppered moths were random the carriers of the *carbonaria* survived since it blended into the darker bark of trees. Carriers of *carbonaria* pass on their genes while most typical moths could not. The change in the color of bark is the selective force which caused the genes to mutate
5	I'm saying letter () because the moths had adapted to their environment, even though the moths were light to stand out, they became darker to survive
6	Because the birds prey range was limited by their ability to see, and because of the darker trees, the whitish moths were easier to see, and therefore more heavily preyed upon by the birds. This led to the darker moths making more offspring, and therefore a change in population
7	I chose () because it described more of what happened than the others. A change in the color of the bark caused the light colored moths to be more visible, a mutation in one of the moths caused it to be black. More black ones survived than white ones because they blended with the habitat better so their amount grew

References

Alexander, R. J. (2008). *Towards dialogic teaching: Rethinking classroom talk*. New York: Dialogos.

Alonzo, A. C., & Gotwals, A. W. (Eds.). (2012). *Learning progressions in science: Current challenges and future directions*. New York: Springer Science & Business Media.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*(3), 389–421.

Atkin, J. M., Coffey, J. E., Moorthy, S., Sato, M., & Thibeault, M. (2005). *Designing everyday assessment in the science classroom*. New York: Teachers College Press.

Ayala, C. C., Shavelson, R. J., Ruiz-Primo, M. A., Brandon, P., Yin, Y., Furtak, E. M., et al. (2008). From formal embedded assessments to reflective lessons: The development of formative assessment suites. *Applied Measurement in Education, 21*(4), 315–334.

Ball, D. L., & Cohen, D. K. (1999). *Developing practice, developing practitioners: Toward a practice-based theory of professional education*. San Francisco: Jossey Bass.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. Retrieved from http://CRAN.R-project.org/package=lme4.

Bell, B., & Cowie, B. (1999). Researching Formative Assessment. In J. Loughran (Ed.), *Researching teaching: Methodologies and practices for understanding pedagogy* (pp. 198–214). London: Falmer Press.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25. doi:10.1080/0969594X.2010.513678.

Bernstein, B. (2000). *Pedagogy, symbolic control, and identity: Theory, research, critique*. Lanham, MA: Rowman & Littlefield Publishers Inc.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74. doi:10.1080/0969595980050102.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8), 3–15.

Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. (2008). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education, 24*(2), 417–436. doi:10.1016/j.tate.2006.11.012.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*(1), 33–64.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications Inc.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and Quasi-experimental designs for research*. Chicago: Rand McNally.

Cazden, C. B. (2001). *The language of teaching and learning*. Portsmouth, NH: Heinemann.

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education, 18*(8), 947–967.

Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching, 48*(10), 1109–1136. doi:10.1002/tea.20440.

Corcoran, T., Mosher, F. A., & Rogat, A. (2009). *Learning progressions in science: An evidence-based approach to reform*. Philadelphia, PA: Consortium for Policy Research in Education.

Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education: Principles, Policy & Practice, 6*(1), 101–116.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualizations and measures. *Educational Researcher, 38*(3), 181–199. doi:10.3102/0013189x08331140.

Dillon, J. T. (1985). Using questions to foil discussions. *Teaching and Teacher Education, 2*, 109–121. doi:10.1016/0742-051X(85)90010-1.

Duschl, R. A., & Gitomer, D. H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment, 4*(1), 37–73.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education, 47*(2), 123–182.

Falk, A. (2012). Teachers learning from professional development in elementary science: Reciprocal relations between formative assessment and pedagogical content knowledge. *Science Education, 96*(2), 265–290. doi:10.1002/sce.20473.

Friedrichsen, P. M., & Dana, Th M. (2003). Using a card-sorting task to elicit and clarify science-teaching orientations. *Journal of Science Teacher Education, 14*(4), 291–309.

Furtak, E. M. (2009a). *Formative assessment for secondary science teachers*. Thousand Oaks, CA: Corwin Press.

Furtak, E. M. (2009). Toward learning progressions as teacher development tools. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning Progressions in Science Conference*. Retrieved from http://education.msu.edu/projects/leaps/proceedings/Default.html.

Furtak, E. M. (2011). *'Flying Blind': An exploration of beginning science teachers' enactment of formative assessment practices*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching, 49*(9), 1181–1210.

Furtak, E. M., & Heredia, S. (2014). Exploring the influence of learning progressions in two teacher communities. *Journal of Research in Science Teaching, 51*(8), 982–1020.

Furtak, E. M., Morrison, D. L., & Kroog, H. (2014). Investigating the link between learning progressions and classroom assessment. *Science Education, 98*(4), 640–673.

Gröschner, A., Seidel, T., Kiemer, K., & Pehmer, A.-K. (2014). Through the lens of teacher professional development components: The "Dialogic Video Cycle" as an innovative program to foster classroom dialogue. *Professional Development in Education*, (September), 1–28.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Heritage, M. (2008). *Learning progressions: Supporting instruction and formative assessment*. Washington, DC: Council of Chief State School Officers.

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*(3), 24–31.

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2014). Addressing knowledge deficits in tutoring and the role of teaching experience: Benefits for learning and summative assessment. *Journal of Educational Psychology, 106*, 934–945.

Jurik, V., Gröschner, A., & Seidel, T. (2013). How student characteristics affect girls' and boys' verbal engagement in physics instruction. *Learning and Instruction, 23*, 33–42. doi:10.1016/j.learninstruc.2012.09.002.

Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Science Education, 98*(4), 674–704.

Kiemer, K., Seidel, T., Gröschner, A., & Pehmer, A.-K. (2015). Effects of a productive classroom discourse intervention on students' motivation to learn mathematics and science. *Learning and Instruction, 35*, 94–103. doi:10.1016/j.learninstruc.2014.10.003.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254.

Kobarg, M., & Seidel, T. (2007). Prozessorientierte Lernbegleitung—Videoanalysen im Physikunterricht der Sekundarstufe I. *Unterrichtswissenschaft, 35*(2), 148–168.

Lee, V. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35*, 125–141.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86–92.

Mehan, H. (1979). *Learning lessons*. Cambridge, MA: Harvard University Press.

Mercer, N. (2010). The analysis of classroom talk: Methods and methodologies. *British Journal of Educational Psychology, 80*(1), 1–14. doi:10.1348/000709909x479853.

Michaels, S., O'Connor, C., & Resnick, L. B. (2008). Deliberative discourse idealized and realized: Accountable talk in the classroom and in civic life. *Studies in Philosophy and Education, 27*(4), 283–297. doi:10.1007/s11217-007-9071-1.

National Research Council. (2001a). *Classroom assessment and the national science education standards*. Washington, DC: National Academy Press.

National Research Council. (2001b). *Inquiry and the national science education standards*. Washington, D.C.: National Academy Press.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. Washington, DC: National Academies Press.

Oliveira, A. W. (2010). Improving teacher questioning in science inquiry discussions through professional development. *Journal of Research in Science Teaching, 47*(4), 422–453. doi:10.1002/tea.20345.

Otero, V. K., & Nathan, M. J. (2008). Preservice elementary teachers' views of their students' prior knowledge of science. *Journal of Research in Science Teaching, 45*(4), 497–523.

Palmer, D. J., Stough, L. M., Burdenski, T. K, Jr, & Gonzales, M. (2005). Identifying teacher expertise: An examination of researchers' decision making. *Educational Psychologist, 40*(1), 13–25.

Penuel, W. R., Fishman, B. J., Cheng, B. H., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. *Educational Researcher, 40*(7), 331–337.

R Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Retrieved from http://www.R-project.org/.

Reznitskaya, A. (2012). Dialogic teaching: Rethinking language use during literature discussions. *The Reading Teacher, 65*(7), 446–456. doi:10.1002/TRTR.01066.

Rimmele, R. (2015). *Videograph,* version 4.2.1.26X3. Leibniz Institute for Science Education, Kiel, Germany.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment, 7*, 99–141.

Ruiz-Primo, M. A., & Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning. *Educational Assessment, 11*(3 & 4), 237–263.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57–84.

Scott, P., Mortimer, E. F., & Aguiar, O. G. (2006). The tension between authoritative and dialogic discourse: A fundamental characteristic of meaning making interactions in high school science lessons. *Science Education, 90*(4), 605–631.

Schoppek, W. (2015). Mehrebenenanalyse oder Varianzanalyse? Ein simulationsbasierter Vergleich von Verfahren zur Auswertung pädagogisch-psychologischer Experimente. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 47*, 199–209.

Seidel, T., & Prenzel, M. (2006). Stability of teaching patterns in physics instruction: Findings from a video study. *Learning and Instruction, 16*(3), 228–240.

Seidel, T., Prenzel, M., Rimmele, R., Herweg, C., Kobarg, M., Schwindt, K., et al. (2007). Science teaching and learning in german physics classrooms—findings from the IPN-video study. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG priority programme* (pp. 79–99). Münster: Waxmann.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 4–14. doi:10.2307/1176145.

Sherin, M. G. (2004). New perspectives on the role of video in teacher education. *Advances in Research on Teaching, 10*, 1–28.

Sherin, M. G., & Han, S. Y. (2002). Teacher learning in the context of a video club. *Teaching and Teacher Education, 20*, 163–183.

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

Smith, J. I., Combs, E. D., Nagami, P. H., Alto, V. M., Goh, H. G., Gourdet, M. A. A., et al. (2013). Development of the biology card sorting task to measure conceptual expertise in biology. *CBE—Life Sciences in Education, 12*, 628–644.

Smith, J. P., DiSessa, A. A., & Roschelle, J. (1993). Misconceptions reconceived: A Constructivist analysis of knowledge in transition. *The Journal of the Learning Sciences, 3*(2), 115–163.

Smith, H., & Higgins, S. (2006). Opening classroom interaction: The importance of feedback. *Cambridge Journal of Education, 36*(4), 485–502.

Snijders, T., & Bosker, R. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Thompson, J., Windschitl, M., & Braaten, M. (2013). Developing a theory of ambitious early-career teacher practice. *American Educational Research Journal, 50*(3), 574–615. doi:10.3102/0002831213476334.

van Driel, J. H., Verloop, N., & de Vos, W. (1998). Developing science teachers' pedagogical content knowledge. *Journal of Research in Science Teaching, 35*(6), 673–695.

Van Veen, K., Zwart, R., & Meirink, J. (2012). What makes teacher professional development effective? A literature review. *Teacher Learning that Matters: International Perspectives*, 3–21.

Whitcomb, J. A. (2013). Learning and pedagogy in initial teacher preparation. In I. B. Weiner (Ed.), *Handbook of psychology* (2nd ed., pp. 441–463). New York: Wiley.

White, R., & Gunstone, R. (1992). *Probing understanding*. New York: Routledge.

Wiliam, D. (2007). Changing classroom practice. *Educational Leadership, 65*(4), 36.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice, 11*(1), 49–65. doi:10.1080/0969594042000208994.

Wilson, S. M. (2013). Professional development for science teachers. *Science, 340*, 310–313. doi:10.1126/science.1230725.

Yin, Y., Shavelson, R. J., Ayala, C. C., Ruiz-Primo, M. A., Brandon, P., Furtak, E. M., et al. (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. *Applied Measurement in Education, 21*(4), 335–359.