

This article was downloaded by: [Stanford University]

On: 27 September 2008

Access details: Access Details: [subscription number 776101540]

Publisher Routledge

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

APPLIED
MEASUREMENT
IN EDUCATION

Applied Measurement in Education

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title-content=t775653631>

On the Impact of Formative Assessment on Student Motivation, Achievement, and Conceptual Change

Yue Yin ^{ab}; Richard J. Shavelson ^c; Carlos C. Ayala ^d; Maria Araceli Ruiz-Primo ^e; Paul R. Brandon ^{1 f}; Erin Marie Furtak ^g; Miki K. Tomita ^{ch}; Donald B. Young ^f

^a College of Education, University of Illinois, Chicago ^b College of Education, University of Hawaii, Manoa ^c School of Education, Stanford University, ^d School of Education, Sonoma State University, ^e University of Colorado at Denver and Health Sciences Center, ^f College of Education, University of Hawaii, ^g Max Planck Institute for Human Development, ^h The Curriculum Research and Development Group, University of Hawaii,

Online Publication Date: 01 October 2008

EA LAWRENCE ERLEBNY ASSOCIATES, PUBLISHERS
Mahwah, New Jersey London

To cite this Article Yin, Yue, Shavelson, Richard J., Ayala, Carlos C., Ruiz-Primo, Maria Araceli, Brandon ¹, Paul R., Furtak, Erin Marie, Tomita, Miki K. and Young, Donald B. (2008) 'On the Impact of Formative Assessment on Student Motivation, Achievement, and Conceptual Change', *Applied Measurement in Education*, 21:4, 335 — 359

To link to this Article: DOI: 10.1080/08957340802347845

URL: <http://dx.doi.org/10.1080/08957340802347845>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

On the Impact of Formative Assessment on Student Motivation, Achievement, and Conceptual Change

Yue Yin

*College of Education
University of Illinois at Chicago
College of Education
University of Hawaii at Manoa*

Richard J. Shavelson

*School of Education
Stanford University*

Carlos C. Ayala

*School of Education
Sonoma State University*

Maria Araceli Ruiz-Primo

University of Colorado at Denver and Health Sciences Center

Paul R. Brandon¹

*College of Education
University of Hawaii*

Erin Marie Furtak

Max Planck Institute for Human Development

¹Order of authors at this point is alphabetical.

Correspondence should be addressed to Yue Yin, Department of Educational Psychology, College of Education, University of Illinois at Chicago, 1040 West Harrison Street, M/C 147, Chicago, IL 60607. E-mail: yueyin@uic.edu

Miki K. Tomita
School of Education
Stanford University
The Curriculum Research and Development Group
University of Hawaii

Donald B. Young
College of Education
University of Hawaii

Formative assessment was hypothesized to have a beneficial impact on students' science achievement and conceptual change, either directly or indirectly by enhancing motivation. We designed and embedded formative assessments within an inquiry science unit. Twelve middle-school science teachers with their students were randomly assigned either to an experimental group ($N = 6$), provided with embedded formative assessment, or control group ($N = 6$). Teachers varied significantly as to their impact on student motivation, achievement, and conceptual change. But the impact of the formative assessment treatment on these outcomes was not statistically significant. Variation in both teachers' classroom management and the degree to which they used informal formative assessment, regardless of group, were conjectured as possible reasons for the absence of an overall formative assessment effect.

The logic of formative assessment—identifying learning goals, assessing where students are with respect to those goals, and using effective teaching strategies to close the gap—is compelling and has led to the expectation that formative assessment would improve students' learning and achievement (Ramaprasad, 1983; Sadler, 1989). This hypothesis has received substantial empirical support (Black & Wiliam, 1998a; Black & Wiliam, 1998b). However, the empirical evidence comes mainly from either laboratory studies or anecdotal records (e.g., Black & Wiliam, 1998a). As Black and Wiliam (1998a) pointed out, studies conducted in laboratory contexts may suffer “ecological validity” problems and encounter reality obstacles when applied in classrooms. The effects of formative assessment have rarely been examined experimentally in regular education settings.

This study examined the effect of formative assessment on student outcomes by using a randomized experiment in a field setting. It explored whether formative assessment would improve student motivation and achievement, and lead to conceptual change. This article presents the (a) conceptual framework, (b) research design, (c) outcome variable measures—motivation, achievement, and conceptual change—and their technical qualities; and (d) results.

CONCEPTUAL FRAMEWORK

We view achievement as acquiring knowledge and the capacity to reason with that knowledge. As described in Shavelson et al. (this issue), knowledge can be distinguished as declarative, procedural, schematic, and strategic. The focus of this study is on schematic knowledge and more specifically, how to move students from a naïve conception of the natural world to a scientifically justifiable one as to “why things sink and float.” Hence we focus on conceptual change and factors that affect it.

Conceptual Change

When entering science classrooms, students often hold deep-rooted prior knowledge or conceptions about the natural world. These conceptions influence how students come to understand what they are taught. Some of their existing prior knowledge provides good foundations for formal schooling, for example, their prior knowledge of number and language. However, other prior conceptions are incompatible with currently accepted scientific knowledge; these conceptions are commonly referred to as *misconceptions* (NRC, 2001) or *alternative conceptions* (Abimbola, 1988), which often arise out of students’ limited observations and experience.

Consequently, learning is not only the acquisition of new knowledge but also the interaction between new knowledge and students’ prior knowledge. For example, everyday life experience leads young children to believe that the earth is like a square. Even when some children learn that the earth is round, they believe that it is like a pancake—round but still flat (Vosniadou & Brewer, 1992), a modified, but unchanged, conception.

The topic of this study, “Why things sink and float,” is addressed in many middle-school physical science curricula and students’ misconceptions about the topic have been well documented (e.g., Hewson, 1986; Kohn, 1993; Yin, Tomita, & Shavelson, 2008). Although sinking and floating is a common experience in everyday life, it is a sophisticated scientific phenomenon. To understand the fundamental reasons underlying sinking and floating, students need to have knowledge that includes an analysis of forces (buoyancy and gravity) and water pressure. That knowledge, however, is neither introduced nor sufficiently addressed in American middle-school curricula probably because of the expected difficulty learning it. Rather, some curriculum developers take a shortcut and use *relative density* to simplify the explanation of sinking and floating (e.g., Pottenger & Young, 1992). Even so, relative density itself is challenging for many students because density is a concept involving the ratio of mass to volume (e.g., Smith, Snir, & Grosslight, 1992) and relative density involves comparing two ratio variables, the density of an object and the density of a medium.

Despite its complexity in science, sinking and floating is a common phenomenon. Most students have rich experiences and personal theories or mental models for explaining sinking and floating. Unfortunately, many of their theories are either misconceptions or conceptions that are only valid under certain circumstances—for example, big/heavy things sink, things with holes in them sink, hollow things float, things with air in them float, or flat things float. Those conceptions are so deeply rooted in students' minds that it is difficult for students to change them, even after students have been intensively exposed to scientific conceptions such as, "If an object's density is less than a liquid's density, the object will float in the liquid regardless of the size or mass of the object" (Yin et al., 2008).

Similar to children's misunderstanding of the shape of the planet earth and "why things sink and float," many other misconceptions are deeply rooted in everyday experiences, widely across different subject domains, among people of different ages, across different cultures, and through the history of the development of scientifically justifiable ideas. These alternative conceptions inhibit students from acquiring scientific conceptions (Yin, 2005).

Given the presence and persistence of alternative conceptions, the restructuring or reorganization of existing knowledge, or *conceptual change*, has become a very important component of teaching and learning (e.g., Carey, 1984; Schnotz, Vosniadou, & Carretero, 1999). To establish scientifically justifiable conceptions of the natural world, sometimes students have to experience conceptual change (Carey, 1984) and transform misconceptions to complete and accurate conceptions (NRC, 2001). From a cognitive perspective, researchers have described the mechanics of conceptual change (Demastes, Good, & Peebles, 1996; Posner, Strike, Hewson, & Gertzog, 1982); developed theories about the nature of conceptual change (Chi, 1992; Chi, Slotta, & deLeeuw, 1994); and explored practical ways to foster conceptual change (Chinn & Malhotra, 2002; Hewson & Hewson, 1984).

Motivation researchers have proposed that conceptual change is influenced by students' motivational beliefs, such as goal orientations, epistemological beliefs, interest or value, and efficacy (e.g., Pintrich, Marx, & Boyle, 1993; Pintrich, 1999). They suggested that conceptual change could be facilitated by students' acquisition of a task goal orientation (the goal of studying is to learn) (Ames, 1992; Dweck & Leggett, 1988; Nicholls, 1984), incremental intelligence belief (intelligence is a malleable quality that can be developed) (Dweck, 1999), high self-efficacy (individuals believe that they have performance capabilities in a particular domain) (Bandura, 1997; Bandura & Cervone, 1983), and strong interest (individuals have positive attitude or preference for the content or task) (Hidi, 1990; Shiefele, Krapp, & Winteler, 1992). In contrast, conceptual change could also be prevented by ego goal orientation (the goal of study is to perform better than others or not worse than others) and fixed intelligence beliefs (intelligence is inborn and fixed) (Dweck, 1999).

Pintrich and his colleagues suggested that *motivational beliefs* influence people's cognitive engagement and depth of information processing, which in turn determines whether conceptual change can occur. Moreover, Pintrich et al. insisted that all motivational beliefs are amenable to change (Pintrich, 1999; Pintrich, Marx, & Boyle, 1993). For instance, an educational context that encourages task goal orientation will encourage students to develop a task goal orientation. Similarly, an ego goal-orientated context encourages ego goal orientation among students. Few studies, however, have been conducted to examine empirically motivation theorists' claims about the influence of motivation on conceptual change. One possible approach to bringing about conceptual change that integrates cognition and motivation is formative assessment.

Formative Assessment

Building on earlier reviews, Black and Wiliam examined 580 articles (or chapters) from over 160 journals in a 9-year period and concluded that formative assessment has a positive impact on students' learning, as well as on students' motivation (Black, Harrison, Lee, Marshall, & Wiliam, 2002; Black & Wiliam, 1998a, 1998b; Hattie & Timperley, 2007).

In summative assessment, scores are often related to a student's rank compared to peers' ranks, and performance differences are the most important concern. This environment may produce several negative effects on motivation and achievement: Students (a) developed an ego-involving goal orientation (Schunk & Swartz, 1993); (b) tended to attribute achievement to capability especially the low achievers (Siero & Van Oudenhoven, 1995) and regarded ability as inborn and intelligence as fixed (Vispoel & Austin, 1995), which might discourage them from putting effort into future learning; (c) were reluctant to seek help because they fear their questions will be regarded as evidence of low ability (Blumenfeld, 1992); and (d) tended to be engaged in superficial information processes, for example, rote learning and concentrating on the recall of isolated details (Butler, 1987; Butler & Neuman, 1995).

Unlike summative assessment, formative assessment is expected to improve motivation and achievement: (a) Formative assessment emphasizes the learning process and closing the gap between students' current situation and the desired goal, so that students may be more likely to form a task-involving goal orientation, which encourages students to process information more deeply than does the performance orientation (Schunk & Swartz, 1993); (b) formative assessment concentrates on improving students' learning, so it may be less likely to cause students to lose confidence. Instead, students tend to believe in incremental intelligence; that is, ability is just a collection of skills that people can master over time (Vispoel and Austin, 1995). Formative assessment shapes self-efficacy in a way that benefits student learning and,

consequently, student achievement, especially that of lower achievers (Geisler-Brenstein & Schmeck, 1996); (c) Formative assessment activities can improve students' interest in learning; and (d) students engaged in formative assessment increase their self-regulation, reasoning, and planning, which are all important for effective learning and conceptual change (Black & William, 1998a).

In sum, formative assessment is expected to encourage the motivational beliefs hypothesized to promote conceptual change, such as task goal orientation, incremental intelligence beliefs, self-efficacy, and interest. Meanwhile, formative assessment discourages the motivational beliefs hypothesized to prevent conceptual change, such as ego goal orientation and fixed intelligence beliefs. For future discussion convenience, the former motivational beliefs are called *positive motivation* and the latter are called *negative motivation*.

Because formative assessment and conceptual change share these motivational beliefs, it seems plausible to connect the motivational model of conceptual change with formative assessment. We conjectured that students who participate in formative assessment would be more likely to change motivational beliefs in a way that promotes conceptual change and achievement than those who did not participate in formative assessment.

As promising as it sounds, formative assessment places greater demands on teachers than regular teaching due to its uncertainty and flexibility (Bell & Cowie, 2001). To date, there is a paucity of systematic, detailed, and practical information to address adequately the challenge of how to design and implement formative assessment in teaching to help improve student learning and motivation.

Shavelson et al. (this issue) characterized formative assessment techniques as falling on a continuum based on the amount of planning involved and the formality of technique used, benchmarking points along the continuum as—(a) on-the-fly formative assessment; (b) planned-for-interaction formative assessment; and (c) formal and embedded in curriculum formative assessment. The first two kinds of formative assessment require teachers to have rich teaching experience and/or strong spontaneous teaching skills. To assist teachers in implementing formative assessment, formal formative assessment, or *embedded formative assessment*, was taken in this study's examination of the impact of assessment on student outcomes.

The study was to explore a systematic way of developing embedded formative assessment (Ayala et al., this issue) and use it to promote conceptual change. This article attempted to answer the following questions: (a) Can embedded formative assessment improve students' motivational beliefs, which were conjectured to influence conceptual change? (b) Can embedded formative assessment improve students' achievement? And (c) Can embedded formative assessment improve conceptual change?

METHOD

Design

To answer the three research questions, a small randomized experiment was conducted. The study involved two groups of teachers, one that employed embedded formative assessment while teaching a science unit and another that taught the same unit without embedded formative assessment. Four steps were taken in the study: (a) Twelve teachers along with their students were randomly assigned to either the experimental group or the control group; (b) All students were pretested on motivation and science achievement (including conceptual change items); (c) Both groups of teachers taught the same curriculum unit provided by the curriculum developer (Pottenger & Young, 1992). Teachers in the experimental group were also provided embedded formative assessment and trained to use the information collected to help improve their teaching and students' learning; and (d) All the students were posttested on motivation, achievement, and conceptual change. By comparing the experimental and control students' scores on the pretest and posttest, we examined whether the embedded formative assessment treatment affected students' motivation, achievement, and conceptual change.

To avoid a Hawthorne effect on experimental teachers or a John Henry effect on control teachers, neither the control nor the experimental teachers were informed about the experiment when they were recruited. Instead, they were told that the study was to assist curriculum designers to improve the curriculum. The control teachers were told to use their regular teaching practice, help to videotape one of their classrooms, and collect data related to student learning. In addition to the data collection requirement for control teachers, the experimental teachers were asked to implement the embedded formative assessments designed by the researchers. Moreover, teachers were in different states, or at least in different schools, so that treatment diffusion was impossible.

Although asked to complete the curriculum unit in half a year, teachers took varying amounts of time, from 63 days to 249 days (Experimental: $M = 130$, $SD = 49$; Control: $M = 106$, $SD = 47$). On average, experimental teachers took 24 days more than the control teachers. This difference was larger than what was expected, because the embedded formative assessments were designed to be completed in 12 more class sessions than in the regular curriculum. The variation in instructional duration might be explained in three ways: first, different teachers' teaching style—some teachers taught much faster than others; second, different school contexts—in some schools, a great proportion of students were English language learners, therefore a large amount of time had to be devoted to English learning including in science class; and third, different state requirements—teachers needed to complete the specific requirements from their states, for example, some teachers needed to prepare their students to pass certain state

tests, therefore, teachers could not completely devote all their teaching time to the curriculum.

Details about the participants and procedures are addressed by Shavelson et al. (this issue). Details about the curriculum and treatment are described by Ayala et al. (this issue). This article focuses on the instruments and results of the study.¹

Instruments

Motivation questionnaire and achievement assessments were developed to examine the impact of embedded formative assessment.

Motivation questionnaire

Motivation questionnaire development. The motivation questionnaire measures the constructs that were addressed in both the conceptual change and formative assessment literatures (Black et al., 2002; Pintrich, 1999; Pintrich et al., 1993). Among them, four motivational beliefs were expected to promote conceptual change: (a) task goal orientation (e.g., “An important reason I do my science work is because I like to learn new things”), (b) perceived task-goal orientation context (e.g., “Our teacher really thinks it is very important to try hard”), (c) self-efficacy in science (e.g., “I am certain I can figure out how to do difficult work in science”), and (d) interest in science (e.g., “I find science interesting”). The other four motivational beliefs were expected to prevent conceptual change: (a) ego approach orientation (e.g., “I want to do better than other students in my science class”), (b) ego avoidance orientation (e.g., “It is very important to me that I do not look stupid in my science class”), (c) perceived performance–goal orientation context (e.g., “Our teacher lets us know which students get the highest scores on tests”), and (d) inborn intelligence epistemic beliefs (e.g., “I can’t change how smart I am in science”).

Most motivation items were drawn from Lau and Roeser’s study (2002). A five-point Likert-type scale from 1 (*strongly disagree*) to 5 (*strongly agree*) was used in the questionnaire to measure the degree to which students agreed or disagreed with each statement. The same motivation questionnaire was given to the control and experimental groups at pretest and posttest.

Technical quality of the motivation questionnaire. Exploratory and confirmatory factor analysis were conducted to examine whether the motivation items measured the constructs that they were designed to measure and to identify ill-matched items that were removed (for details see Yin, 2005). The results

¹Due to space limitations, the discussion of instruments and results in this article is rather brief. To know more details about the instruments and results, please refer to Yin (2005).

based on pretest data, similar to the result based on posttest reported, were omitted to conserve space.

Overall, the motivation sub-scales had acceptable internal consistency except for fixed intelligence, with an alpha coefficient of .44 due to small number of items ($N = 3$). Alpha coefficients for the rest of the motivation sub-scales were .83 for task goal orientation (item number $N = 5$), .86 for perceived task goal orientation ($N = 6$), .89 for self-efficacy ($N = 5$), .81 for interest ($N = 3$), .77 for ego approach goal orientation ($N = 4$), .78 for ego avoidance goal orientation ($N = 5$), and .70 for perceived performance goal orientation ($N = 6$). A composite score for each construct was calculated by averaging the item scores in each construct. Eight composite scores were used for further data analyses.

All the positive motivation constructs were positively correlated with each other, with disattenuated correlations ranging from .64 to .85, $p < .01$ ($N = 788-826$). Similarly, all negative motivation constructs were positively correlated with each other, with disattenuated correlations ranging from .37 to .84, $p < .01$ ($N = 788-826$). Except that ego-approach and self-efficacy were positively correlated with low magnitude ($r = .10$), positive motivation constructs were either uncorrelated or negatively correlated with negative motivation constructs, ranging from .00 to $-.53$.

Exploratory factor analysis, on the motivation subscale composite scores, with Principal axis factoring and Promax rotation, further confirmed the correlation results. The positive motivation constructs loaded mainly on the first factor with loadings ranging from .73 to .84, whereas the negative motivation constructs were loaded mainly on the second factor with loadings ranging from .47 to .76. Both the correlation pattern among the motivation constructs and principal component analysis provided evidence for construct validity of the motivation measurement.

Achievement assessments

In this section, the multiple-choice test and open-ended assessments are described and their technical qualities examined.

Achievement assessment development. Assessment development focused on: (a) FAST instructional objectives and (b) representation of the different knowledge types—in particular, declarative, procedural, and schematic knowledge (see Shavelson et al., this issue). Four assessments were developed: a multiple-choice test, a performance assessment, a short-answer assessment, and a Predict-Observe-Explain assessment. A description of the Predict-Observe-Explain assessment can be found in Ayala et al. (this issue).

Thirty-eight items were included in the multiple-choice pretest, covering the important instructional objectives across three types of knowledge. Multiple-choice

test item examples can be found in Shavelson et al. (this issue). In addition to the 38 items on the pretest, five more items mainly focusing on common misconceptions about sinking and floating were included on the posttest. Among multiple choice items, 14 items were adopted from external sources, such as the National Assessment of Educational Progress and Trends of International Mathematics and Science Study, to tap the spread of the treatment effect. Because the correlation between the external items and internal items was high, $r = .70$, $p < .01$, the external items were analyzed with the internal items together in this article.

A performance assessment was designed to assess students' procedural and schematic knowledge. In the first part of the performance assessment, students were given information about the mass of a rectangular block and some equipment, and asked to find the solid block's density. To solve the problem, students needed to (a) use either a ruler or an overflow container to measure the volume of the block and (b) apply the density formula. In the second part of the performance assessment, students were given three blocks with density information and a bottle of mystery liquid with unknown density. They were asked to find the density of the mystery liquid. To solve this problem, students needed to conduct an experiment, observe the sinking and floating behavior of three blocks in the mystery liquid, and infer the mystery liquid's density.

The short-answer assessment was designed to measure students' declarative and schematic knowledge. It asked students to explain "why do things sink or float" and to provide evidence and an example to support their argument.

The Predict-Observe-Explain assessment was also designed to assess students' schematic knowledge. The Predict-Observe-Explain assessment in the posttest and those used in the embedded formative assessments were similar in format, but different in content. The soap Predict-Observe-Explain on the posttest was intended to examine whether students understood two main points: (a) density is a property of a material and will not change with size and (b) an object sinks or floats depending on its density (relative to the medium's density) instead of its volume and mass.

The multiple-choice test was administered at both pretest and posttest. Performance, short-answer, and Predict-Observe-Explain assessments were administered at posttest only because they are heavily content loaded and costly to administer. Research team members administered the instruments in the classrooms of each teacher immediately before and after the target unit. The motivation questionnaire was given before the achievement tests, so that students' report of their motivation would not be influenced by their performance on the achievement tests. On the posttest, the achievement tests were given in the following order: multiple-choice, performance assessment, short-answer assessment, and Predict-Observe-Explain assessment. The Predict-Observe-Explain task was given last because of its possible instructional effect.

Technical quality of the achievement assessments. The quality of the multiple-choice test items was examined as to item difficulties and instructional sensitivity. The quality of the three open-ended assessments was examined as to inter-rater reliability/agreement. In addition, the internal consistency and the construct validity of all the assessments were examined.

With respect to difficulty, most items appropriately fell in the moderate range (p -values .40–.80 at posttest). Instructional sensitivity (D) was used to examine whether multiple-choice test items discriminated between examinees who had and those who had not been taught (Crocker & Algina, 1986). All the items had positive instruction sensitivity D s, indicating that more students chose correct answers at posttest than at pretest. This result provided evidence for the content validity of the multiple-choice test items. That is, the items were linked to the curriculum content and would be expected to reflect the effect of instruction.

In order to capture information about students' achievement reflected in their responses to the assessments, analytical scoring systems were developed for each assessment (Yin, 2005). Doctoral students in science education scored the three open-ended assessments: performance assessment (six raters), short-answer assessment (six raters), and Predict-Observe-Explain assessment (four raters). Before independent scoring, all raters received scoring training until they reached satisfactory inter-rater reliability (above .80) or agreement (above 80%). Performance assessments were scored with numerical values assigned by six raters. Therefore, G-coefficients were used to estimate reliability. Based on 48 randomly selected student responses on the performance assessment, the average G-coefficient for the six raters (three in each group) was .83. Short-answer and Predict-Observe-Explain assessments were scored with categorical values, therefore agreements were used to evaluate the inter-rater agreement. Based on 48 randomly selected student responses on short-answer, the average agreement of nine raters (two, three, and four raters in each group) was 87%. Based on 56 randomly selected student responses on Predict-Observe-Explain, the four raters' agreement was 92%.

After all the assessments were scored, the internal consistency of each assessment and the correlations among the assessments were calculated. Alpha coefficients were .86 for multiple-choice test (item number $N = 43$), .81 for performance assessment ($N = 18$), .83 for short-answer ($N = 4$), and .74 for Predict-Observe-Explain ($N = 7$), indicating acceptable internal consistencies. Because the four assessments measure students' knowledge about the same topic but with different emphases, high internal consistency within each assessment and moderately high correlations among assessments were expected. Correlations between different assessments ranged from .39 to .69, confirming our expectation and providing evidence for construct validity.

Conceptual change score. Fourteen selected items from the four achievement tests focused on conceptual change. Internal consistency of the 14 items

was .79. A composite score was calculated and interpreted as conceptual change, with a maximum of 14. In addition to the score on each assessment, the conceptual change composite score as well as achievement total score were used in the analyses that follow.

Data Analysis Plan

Because teachers were in different states or different school districts in a state, the teaching contexts varied dramatically across teachers, such as the number and size of classes a teacher taught. Moreover, student achievement and motivation subscale scores varied greatly across classes and teachers at pretest, even though pairs of teachers were matched on school-level variables and randomly assigned to the treatment or control group based on school size, the percent of reduced lunch, and math/reading proficiency rate (Shavelson et al., this issue). To reduce the initial differences between two groups, one class, called *focus class*, was selected for intensive study from each teacher based on students' mean pretest motivation and achievement scores such that the pairs of classes selected were as close as possible.

This article reports on the formative assessment impact on focus classes for three reasons: (a) The focus classes were the most closely matched classes of teachers in the control and experimental groups; (b) the focus classes were the classes from which we collected the most elaborate implementation information; and (c) the answers to the research questions based on focus classes were similar to those based on all the students.

Although the focus classes were the closest matches, significant differences existed among some teacher pairs (Table 1). Significant differences exist among three teacher pairs on motivation scores. Two experimental teachers' students had significantly higher positive motivation and lower negative motivation than their counterparts in the control group. One control group teacher's students had significantly lower negative motivation score than their counterparts in the experimental group. Imbalance also existed on achievement. Students in the control group of two teachers scored significantly higher on the multiple-choice test than did their experimental group counterparts on average (Yin, 2005). Overall, the experimental group started with higher average motivation scores but lower mean achievement scores than the control group at pretest. Although not completely adequate, covariate adjustments were made in statistical analyses.

To answer the research questions, the control and experimental groups were first compared descriptively on mean scores at posttest. Because students were nested in teachers and teachers were nested in treatment groups, hierarchical linear modeling (HLM) was then applied to examine the formative assessment treatment's influence statistically. It should be noted that in

TABLE 1
Comparison of Experimental and Control Group Student Motivation, Achievement, and Conceptual Change Scores at Pretest

Group Pretest Measures	Experimental			Control			Difference
	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>t</i>
Task Goal	144	3.88	.61	133	3.72	.66	2.12*
Task Perceived	144	4.20	.47	133	4.13	.48	1.38
Self-Efficacy	144	4.04	.55	133	3.91	.58	1.99*
Interest	144	4.14	.71	133	3.96	.76	2.06*
Ego Approach	144	2.95	.86	133	3.07	.82	-1.14
Ego Avoidance	144	2.97	.85	133	3.08	.80	-1.04
Performance Perceived	144	2.24	.60	133	2.51	.60	-3.82**
Fixed Ability	144	2.39	.84	133	2.27	.76	1.19
Multiple-choice	145	14.22	4.30	135	15.39	4.49	-2.22*

Note. * $p < .05$; ** $p < .01$.

this study HLM lacks statistical power due to the small number of teachers in each group.²

RESULTS

Preliminary Analysis of Pretest and Posttest Scores

Due to the nested structure, it is not statistically legitimate to combine all the students in each treatment group for data analysis. However, the descriptive analyses based on the combined groups can provide a preliminary overall picture of the two group comparison and treatment effect at posttest. Information in Table 1 indicates that overall the experimental group scored higher than the control group on most positive motivation scores but scored lower on achievement at pretest.

Experimental group still scored higher than control group on positive motivation scales at posttest (Table 2), but only the mean difference in perceived task goal orientation was statistically significant. As for achievement, the control group outperformed the experimental group, scoring significantly higher, on average, on the multiple-choice test, performance assessment, and total achievement score. Based on the comparison of Table 1 and Table 2, the treatment

²The intent of the study was exploratory and we and the National Science Foundation agreed on the small sample size for cost reasons and as a first step in providing an "existence proof" that might lead to future funding.

TABLE 2
Comparison of Experimental and Control Group Student Motivation, Achievement, and Conceptual Change Scores at Posttest

Group Posttest Measures	Experimental (E)			Control (C)			Difference
	N	Mean	SD	N	Mean	SD	t
Task Goal	125	3.55	.79	123	3.53	.84	0.14
Task Perceived	125	4.14	.62	123	3.95	.71	2.25*
Self-Efficacy	125	3.97	.62	123	3.87	.65	1.25
Interest	125	3.76	1.00	123	3.59	1.12	1.29
Ego Approach	125	2.85	.92	123	2.99	.86	-1.24
Ego Avoidance	125	2.79	.91	123	2.72	.88	0.57
Performance Perceived	125	2.29	.72	123	2.31	.67	-0.25
Fixed Ability	125	2.14	.72	123	2.25	.78	-1.14
Multiple-choice	129	22.92	6.59	125	25.15	7.16	-2.58**
Performance Assessment	117	13.05	7.11	123	15.31	6.99	-2.48*
Predict-Observe-Explain	120	2.53	1.93	119	2.74	2.26	-0.79
Short-Answer	125	2.24	1.49	128	2.37	1.53	-0.67
Achievement Total	107	41.55	13.94	107	46.41	15.31	-2.43*
Conceptual Change	107	5.32	2.99	107	5.44	3.52	-0.29
Multiple-Choice Gain	127	8.82	5.83	120	9.58	6.49	-0.96

Note. * $p < .05$; ** $p < .01$.

seemed not to have a statistically significant impact on student motivation, achievement or conceptual change at posttest.

However, in general, experimental group students had significantly lower score variance than the control group on the Predict-Observe-Explain assessment ($F = 4.09$, $p < .05$) and conceptual change score ($F = 3.88$, $p = .05$), that is, the achievement gap between higher achievers and lower achievers in the experimental group was not as big as that in the control group. This is consistent with the finding that formative assessment is particularly useful for low achievers (Black & Wiliam, 1998a).

HIERARCHICAL LINEAR MODELING

Variables Used in the HLM

Posttest motivation, achievement, and conceptual change scores served as dependent variables in the multivariate model. For analysis efficiency, four positive motivation subscales, four negative motivation subscales, and four achievement tests were analyzed together, respectively. A three-level HLM model was specified with multiple outcomes nested in students at level-1,

students nested within classes at level-2, and teachers at level three. Because the achievement tests used different scales, achievement scores were standardized before the analysis.

Independent variables, or predictors, include pretest scores and treatment group. For analysis efficiency and statistical power, at level-2 positive motivation and negative motivation factor scores were respectively used as predictors, rather than individual motivation constructs, in the motivation analyses. Pretest multiple-choice scores were used as predictor at level-2 in the achievement analysis. Treatment group was the predictor at level-3. Due to heterogeneous school context and teacher backgrounds, many teacher and school level predictors exist at level-3, such as, school size, ethnic composition, the percent of students who receive free or reduced price lunch, students' grade level, teachers' teaching experience, highest educational degree, science background, teaching load, and length of instruction. However, due to the constraint of small sample size (12 teachers at level-3), only the treatment group, the focus in this study, was introduced into the model as the predictor as level-3.

Different from the analysis of motivation and achievement, conceptual change was measured by one variable only, therefore two levels were involved in the HLM: at level-1, conceptual change was the outcome variable, pretest multiple-choice score was the predictor; at level-2, treatment group was the predictor.

Model Construction for Achievement and Motivation Measures

Three steps were taken to analyze achievement and motivation variables:

Step 1: Unconditional model

An unconditional model examined the variation in all of the outcome variables at three levels: multiple outcome variables at level-1, student at level-2, and teacher at level-3 (1 to 3).

Level-1

$$Y = \pi_0 + e \tag{1}$$

Level-2

$$\pi_0 = \beta_{00} + r_0 \tag{2}$$

Level-3

$$\beta_{00} = r_{000} + \mu_{00} \tag{3}$$

Y was the outcome variable, such as multiple positive motivation scores, negative motivation scores, or achievement scores. π_0 was the intercept, the mean score for each student. e was the random error. β_{00} was the intercept, in this case, the mean score for each teacher's students. r_0 was the random error at student level. Variance due to r_0 represented that the score variation at level-2, in this case, among students within teachers. β_{00} was a function of grand mean, γ_{000} , plus a random error, μ_{00} . Significant variance, τ_{00} , due to μ_{00} indicated that teacher mean score varied at level-3.

Step 2: Conditional model with covariate(s) at level-2

According to Raudenbush and Bryk (2002), statistical adjustments for individuals' background are important, because (a) "persons are not usually assigned at random to organizations, failure to control for background may bias the estimates of organization effect" (p. 111) and (b) if predictors (or covariates) are strongly related to the outcome of interest, controlling for them will increase the precision of any estimates of organization effects and the power of hypothesis tests by reducing unexplained error variance.

In this study, students differed in mean pretest motivation and achievement scores; therefore, the pretest scores were added to the model as covariate(s) at level-2 to account for the outcome variable variance within-teacher. Only the pretest score corresponding to the outcome score was introduced in the model at level-2. For example, when the posttest positive motivation subscales were the outcome variables, only the pretest positive motivation factor score was introduced as the predictor at level-2. When posttest achievement scores were the outcome variables, only the pretest multiple-choice test score was introduced as the predictor at level-2. Only one covariate was used at level-2 for two reasons: First, the corresponding pretest variable was the most relevant covariate of the corresponding posttest score. Second, other variables were found to be non-significant predictors, for example, no significant relationship was found between achievement and the positive and negative motivation measures. Because pretest scores were continuous variables, they were grand-mean centered (Raudenbush & Bryk, 2002).

The model was specified as follows.

Level-1

$$Y = \pi_0 + e \quad (4)$$

Level-2

$$\pi_0 = \beta_{00} + \beta_{01}(X - \bar{X}) + r_0 \quad (5)$$

Level-3

$$\beta_{00} = \gamma_{000} + \mu_{00} \tag{6}$$

$$\beta_{01} = \gamma_{010} + \mu_{01} \tag{7}$$

In this model, X was a pretest score as the predictor for the corresponding posttest score. The intercept, β_{00} , was the expected outcome for a subject whose value on X was equal to the grand mean, \bar{X} . The slope, β_{01} , represented the regression parameter between students' pretest and posttest scores. r_0 was the random error after controlling for students' pretest score. β_{00} and β_{01} might vary across teachers in the level-3 model as a function of a grand mean and a random error. Because X was grand-mean centered, γ_{000} was the adjusted mean of the teachers on students' posttest scores. γ_{010} was the adjusted mean of pretest and posttest regression slope across those teachers. μ_{00} was the unique increase to the intercept associated with a certain teacher. Significant variance associated with μ_{00} indicated that the mean posttest score varied across teachers. μ_{01} was the unique increase to the slope associated with a certain teacher.

Analysis showed that variance of μ_{01} was not significantly different from 0. That is, the regression coefficients between outcome variable and predictor variable within teachers at level-1 did not vary across teachers at level-3 ("homogeneity of regression slopes"). Therefore, μ_{01} was eliminated from the final model. Accordingly, Equation 7 was simplified into 8.

$$\beta_{01} = \gamma_{010} \tag{8}$$

Step 3: Model with intercepts-as-outcomes

One of the important purposes for using HLM is to examine the cross level effects. As the regression coefficients at level-2 did not significantly vary across teachers, that is, homogeneity of regressions slopes, only the intercept at level-2 was an outcome variable at level-3. As mentioned earlier, treatment group was the only predictor at level-3. Because treatment group was a dummy variable (experimental group = 1, control group = 0), it was added to the model without being centered (Raudenbush & Bryk, 2002).

Analyses showed that the variance of μ_{00} in Equation 6 significantly differed from 0. That is, the intercept at level-2 varied significantly at level-3 across teachers. A predictor at level-3, treatment group, was then included in the model to account for the intercept variance at level-3. The model was specified in Equation 9.

$$\beta_{00} = \gamma_{000} + \gamma_{001}W + \mu_{00} \tag{9}$$

W was the treatment variable at level-3. Treatment group was coded as “0” for control group and “1” for experimental group.

Model Construction for Conceptual Change

Because only one score measured conceptual change, a two-level HLM was used in its analysis. 10 to 12 present the final model used: at level-1, conceptual change, *Y*, was the outcome variable; pretest achievement, *X*, was the covariate. At level-2, the treatment variable, *W*, was the predictor. Again, motivation measures were not included at level one because they were not significant predictors for conceptual change, unlike what was hypothesized in literature (Pintrich, 1999).

Level-1

$$Y = \beta_0 + \beta_1(X - \bar{X}) + r \tag{10}$$

Level-2

$$\beta_0 = \gamma_{00} + \gamma_{01}W + \mu_0 \tag{11}$$

$$\beta_1 = \gamma_{10} \tag{12}$$

HLM Results

Table 3 presents the results when pretest scores were included as covariates and treatment was included as a predictor, the full model for each analysis. The

TABLE 3
Multilevel Regression Estimates for Posttest Scores

<i>Outcome Variable (Posttest)</i>	<i>Predictors</i>	<i>N</i>	<i>Coefficient</i>	<i>t</i>
Positive Motivation	Level-2 Pre Positive Motivation, β_{01}	239	0.28	6.65**
	Level-3 Group, γ_{001}	12	0.07	0.40
Negative Motivation	Level-2 Pre Negative Motivation, β_{01}	239	0.30	9.48**
	Level-3 Group, γ_{001}	12	0.04	0.43
Achievement	Level-2 Pre Achievement, β_{01}	256	0.33	8.69**
	Level-3 Group, γ_{001}	12	-0.11	-0.50
Conceptual Change	Level-1 Pre Achievement, β_1	288	0.28	6.55**
	Level-2 Group, γ_{01}	12	0.01	0.02

Note. ***p* < .01.

results indicate that all the pretest scores were significant positive predictors for their corresponding posttest scores and the inclusion of pretest scores reduced the variance within teachers on motivation and achievement outcomes but not for conceptual change. The positive coefficients for treatment in the positive motivation and conceptual change analyses indicate slight adjusted mean advantage for the experimental group. On the other hand, positive treatment coefficient in the negative motivation analysis and negative treatment coefficient in achievement analysis indicate slight adjusted mean advantage for the control group. However, treatment was not a statistically significant predictor of any outcome—that is, treatment effects did not explain the variance across teachers. These results are consistent with descriptive statistics.

Variance components for motivation scores, achievement scores, and the conceptual change score show that each of these scores varied significantly across teachers. Even after students' pretest scores were controlled at level-2 and treatment group introduced at level-3, variance associated with teachers at level-3 was significant for positive motivation (0.07, $p < .01$), negative motivation (0.01, $p < .05$), achievement (0.12, $p < .01$), and conceptual change (0.32, $p < .01$)

To summarize, based on the HLM analyses, students' motivation, achievement, and conceptual change scores significantly varied among students and across teachers. However, the embedded formative assessment treatment did not explain the variation among teachers. Experimental group students did not seem to benefit, on average, from the embedded formative assessment they received.

DISCUSSION AND CONCLUSION

In this study, the impact of embedded formative assessment on measures of students' motivation, achievement, and conceptual change was explored. Empirical evidence was provided in support of the reliability and validity of scores from these measures.

Both the descriptive statistics and HLM analyses showed that the assessments embedded in the curriculum used by the experimental group *did not* have a significant influence, on average, on students' motivation, achievement, and conceptual change compared to students in the control group that used the curriculum without embedded assessments. This said, regardless of treatment group, teachers varied greatly on the outcomes they produced in their students. This finding did not support our conjectures about the salutatory effect of formative assessment on student outcomes or the findings reported in literature reviews. More elaborate analyses showed that teachers' educational background, teaching experience, teaching load, instruction length, and some school information (such as school size, and percent free and reduced lunch) were not significant predictors

for students' motivation, achievement, and conceptual change difference either (Yin, 2005).

Preliminary analysis of teachers' classroom instruction suggested that students had higher achievement and were more likely to change conceptions if a teacher had the following characteristics: (a) good classroom management, so students in their classes could concentrate on learning and benefit from instruction; (b) successful teaching strategies; and most importantly (c) effective formative-assessment implementation, either formally or informally (see also Furtak et al., this issue). Some experimental teachers did not implement the formal formative assessment as designed, whereas some control teachers implemented informal formative assessment in class and provided timely feedback to students spontaneously when needed, although they were not provided the formal formative assessment as a tool. That might be one of the most important reasons formative assessment did not work as hypothesized (Yin, 2005).

In this study, embedded formative assessment did not have the impact expected on students' motivation, achievement, or conceptual change. However, this result did not disconfirm the effectiveness of formative assessment. Rather, it provided evidence for the difficulty and importance of effectively implementing formative assessment. Simply embedding assessments in curriculum will not impact students' learning and motivation, unless teachers use the information from embedded assessment to modify their teaching. Furtak et al. (this issue) discuss implementation issues in detail.

REFLECTIONS AND FUTURE DIRECTIONS

Experimental Study

The study was planned as an experimental study. Based on information available during the research design phase, the participating teachers were matched pairwise and randomly assigned to experimental and control groups. Despite careful planning, it is much more difficult to conduct a "perfect" experimental study in education than in natural sciences or a psychology laboratory, because it is almost impossible to control for many factors. For example, a school's environment cannot be adequately measured and controlled like the temperature and pressure in a physics lab, and teachers' teaching time cannot be controlled like the reaction time in a chemistry experiment. Consequently, many extraneous variables existed in the study without adequate control, such as, community environment, school environment, teachers' teaching experience, educational background, teaching load, teaching schedule, and students' family background and academic preparation. These variations cannot be measured easily by simple indicators, such as percentage of free or reduced price lunch. Moreover, the small

number of teachers in each group could not balance out the noise for the experimental study. Researchers should pay more attention to get compatible experimental and control groups or increase sample size in future experimental studies.

On Experimental Teacher Training

Another possible reason for the absence of formative assessment effects was the inadequate implementation of the treatment by the experimental teachers, even after training. Some experimental-group teachers were not used to implementing formative assessment or they misunderstood the purpose of using it. For example, experimental teacher Robert seemed to misunderstand the role of formative assessment in an inquiry-based curriculum—he asked students questions but withheld correct answers. According to a researcher who visited Robert's classroom, Robert taught more effectively in a class where he was not required to use embedded assessments. That is, instead of being helped, Robert's teaching might have been "hurt" by the embedded assessment interventions. Some teachers considered using embedded formative assessment as an obligation to Stanford. On the classroom video, experimental teacher Aden inserted students' responses to the formative assessment in an envelope, sealed it, and told his students, "It will be mailed to Stanford." Aden seemed to only use the formative assessment to collect data "for the researchers," but did not use it to help teaching and learning.

To help teachers to effectively implement formative assessment, the following actions might be taken. First, teachers could be invited to participate in the development of formative assessment. If the participant teachers are given more opportunities to participate in assessment design, the assessments could be tailored to teachers' needs and preference and teachers might also feel more comfortable using *their own* assessments. After our study was conducted, Wiliam et al. (2004) published their study, in which teachers developed their own formative assessment strategies. Their study led to encouraging results.

Second, teachers should receive follow up, in-progress coaching from researchers in addition to the intensive workshop training. Weekly phone calls between teachers and researchers in our study did not fulfill this function. Classroom videotapes revealed that teachers also needed formative assessment and feedback on their own teaching practices in order to learn how to use formative assessment. Some teachers struggled with using the embedded formative assessment. Unfortunately, we did not give teachers timely feedback because of their tight schedules and geographical distance on the one hand, and our concern with interfering with the experimental treatment on the other. In the future, researchers should closely monitor teachers to implement the treatment to ensure the fidelity of the treatment.

Third, more research should be done on how to help teachers give effective feedback to students. The embedded formative assessment used in this study succeeded in eliciting students' conceptions; however, drawing on students' conceptions, some teachers were not able to provide helpful feedback so as to close the gap between students' current and desired levels of achievement. During teaching, one teacher in this study sent an e-mail to the researchers asking how she should respond to students who believed that an object with air in it floats. Classroom videotapes showed that this teacher was not the only teacher having difficulty correcting students' deep-rooted alternative conceptions. Strategies for dealing with students' responses could be suggested to or explored with teachers. For example, when students are not using scientific terms, what questions might teachers ask to draw attention to correct terms and the importance of their use? Similarly, when students have alternative conceptions of sinking and floating, what activities could be conducted to help students see the discrepancies between their understanding and the scientifically accepted understanding, gradually moving them towards established scientific conceptions?

Finally, in training teachers, extensive use of video examples might be made to help teachers develop an intuitive understanding of how to use formative assessment. Although researchers and a model teacher demonstrated how to implement embedded formative assessment in the five-day training, those examples seemed insufficient for teachers to learn how to implement embedded formative assessment in the designed way. More videotape examples might have helped teachers get a more and better sense of it. For example, some teachers in both the experimental and control groups successfully used formative assessment (formally or informally). Scenarios of their teaching videos might be used to develop vignettes demonstrating exemplary use of formative assessment in the future studies. These vignettes could then be used to help prepare future teachers to implement formative assessment. As Black and Wiliam suggested, "they [teachers] need to see examples of what doing better means in practice (Black & Wiliam, 1998b, p. 146)."

Although formative assessment is promising instructional technique, it is not a magic bullet. Simply embedding formative assessment in curriculum does not guarantee improved learning and teaching. Teachers need tremendous support using assessment in their teaching practice. Moreover, teachers must also figure out how best to adapt formative assessment to their needs and the need of their students. As pointed out by Black and Wiliam, ". . . if the substantial rewards promised by the evidence are to be secured, each teacher must find his or her own patterns of classroom work. Even with optimum training and support, such a process will take time" (1998b, p. 147).

ACKNOWLEDGMENTS

This research was supported, in part, by a grant from the National Science Foundation (NSF/Award ESI-0095520) and, in part, by the National Center for Research on Evaluation, Standards, and Testing (CRESST/Award 0070 G CC908-A-10).

REFERENCES

- Abimbola, I. O. (1988). The problem of terminology in the study of student conceptions in science. *Science Education*, 72(2), 175–184.
- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84, 261–271.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Bandura, A., & Cervone, D. (1983). Self-evaluative and self-efficacy mechanisms governing the motivational effects of goal systems. *Journal of Personality and Social Psychology*, 45(5), 1017–1028.
- Bell, B., & Cowie, B. (2001). *Formative assessment and science education*. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). *Testing, motivation and learning*. Cambridge: University of Cambridge, Faculty of Education: The Assessment Reform Group.
- Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education*, 5, 7–68.
- Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80(2), 139–148.
- Blumenfeld, P. C. (1992). Classroom learning and motivation: Clarifying and expanding goal theory. *Journal of Educational Psychology*, 84(3), 272–281.
- Butler, R. (1987). Task-involving and ego-involving properties of evaluation. *Journal of Educational Psychology*, 79, 474–482.
- Butler, R., & Neuman, O. (1995). Effects of task and ego achievement goals on help-seeking behaviors and attitudes. *Journal of Educational Psychology*, 87, 261–271.
- Carey, S. (1984). *Conceptual change in childhood*. Cambridge, MA: MIT Press.
- Chi, M. T. H. (1992). Conceptual change within and across ontological categories: Examples from learning and discovery in science. In R. Giere (Ed.), *Cognitive models of science: Minnesota studies in the philosophy of science* (pp. 129–160). Minneapolis: University of Minnesota Press.
- Chi, M. T. H., Slotta, J. D., & deLeeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. *Learning and Instruction*, 4, 27–43.
- Chinn, C. A., & Malhotra, B. A. (2002). Children's responses to Anomalous Scientific Data: How is conceptual change impeded? *Journal of Educational Psychology*, 94(2), 327–343.
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace, PA: Jovanovich College Publishers.
- Demastes, S. S., Good, R. G., & Peebles, P. (1996). Patterns of conceptual change in evolution. *Journal of Research in Science Teaching*, 33, 407–431.
- Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality and development*. Philadelphia: The Psychology Press.
- Dweck, C. S., & Leggett, E. L. (1988). A Social-cognitive approach to motivation and personality. *Psychological Review*, 95(2), 256–273.

- Geisler-Brenstein, E., & Schmeck, R. R. (1996). The revised inventory of learning processes: A multifaceted perspective on individual differences in learning. In M. Birenbaum & F. J. R. C. Dochy (Eds.), *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 284–317). Boston, MA: Kluwer Academic Publishers.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112.
- Hewson, M. G. (1986). The acquisition of scientific knowledge: Analysis and representation of student conceptions concerning density. *Science Education*, 70(2), 159–170.
- Hewson, P. W., & Hewson, M. G. (1984). The role of conceptual conflict in conceptual change and the design of science instruction. *Instructional Science*, 13(1), 1–13.
- Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research*, 60, 549–571.
- Kohn, A. S. (1993). Preschoolers' reasoning about density: Will it float? *Child Development*, 64, 1637–1650.
- Lau, S., & Roeser, R. W. (2002). Cognitive abilities and motivational processes in high school students' situational engagement and achievement in science. *Educational Assessment*, 8(2), 139–162.
- Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review*, 91, 328–346.
- NRC. (2001). *Knowing what students know*. Washington, DC: National Academies Press, National Research Council.
- Pintrich, P. R. (1999). Motivational beliefs as resources for and constraints on conceptual change. In W. Schnotz, S. Vosniadou & M. Carretero (Eds.), *New perspectives in conceptual change research* (pp. 33–50). Oxford, England: Pergamon Press.
- Pintrich, P. R., Marx, R., & Boyle, R. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research*, 63, 167–199.
- Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. F. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66, 211–227.
- Pottenger, F. M. I., & Young, D. B. (1992). *The local environment: FAST 1, foundational approaches in science teaching*. Honolulu: University of Hawaii Curriculum Research & Development group.
- Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, 28(1), 4–13.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.
- Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119–144.
- Schnotz, W., Vosniadou, S., & Carretero, M. (Eds.). (1999). *New perspectives in conceptual change research*. Oxford, England: Pergamon Press.
- Schunk, D. H., & Swartz, C. W. (1993). Goals and progress feedback: Effects on self-efficacy and writing achievement. *Contemporary Educational Psychology*, 18, 337–354.
- Shiefele, U., Krapp, A., & Winteler, A. (1992). Interest as a predictor of academic achievement: A meta-analysis of research. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 183–211). Hillsdale, NJ: Erlbaum.
- Siero, F., & Van Oudenhoven, J. P. (1995). The effects of contingent feedback on perceived control and performance. *European Journal of Psychology of Education*, 10, 13–24.
- Smith, C., Snir, J., & Grosslight, L. (1992). Using conceptual models to facilitate conceptual change: The case of weight-density differentiation. *Cognition and Instruction*, 9(3), 221–283.
- Vispoel, W. P., & Austin, J. R. (1995). Success and failure in junior high school: A critical incident approach to understanding students' attributional beliefs. *American Educational Research Journal*, 32(2), 377–412.

- Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, 24, 535–585.
- William, D., Lee, C., Harrison, C., & Black, P. J. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy and Practice*, 11(1), 49–65.
- Yin, Y. (2005). *The influence of formative assessments on student motivation, achievement, and conceptual change*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.
- Yin, Y., Tomita, K. M., & Shavelson, R. J. (2008). Diagnosing and dealing with student misconceptions about “Sinking and Floating.” *Science Scope*, 31(8), 34–39.