Assessing Science Learning

Perspectives From Research and Practice

Edited by Janet Coffey, Rowena Douglas, and Carole Stearns

NSTA press
National Science Teachers Association

Assessing Science Learning

Perspectives From Research and Practice

Assessing Science Learning

Perspectives From Research and Practice

Edited by Janet Coffey, Rowena Douglas, and Carole Stearns

NSTA press
National Science Teachers Association
Arlington, VA

NSTApress®
National Science Teachers Association

Claire Reinburg, Director
Jennifer Horak, Managing Editor
Judy Cusick, Senior Editor
Andrew Cocke, Associate Editor
Betty Smith, Associate Editor

Art and Design
Will Thomas, Jr., Director

Printing and Production
Catherine Lorrain, Director
Nguyet Tran, Assistant Production Manager

National Science Teachers Association
Gerald F. Wheeler, Executive Director
David Beacom, Publisher

11 10 09 08 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Assessing science learning : perspectives from research and practice / edited by Janet E. Coffey, Rowena Douglas, and Carole Stearns.
p. cm.
Includes index.
ISBN 978-1-93353-140-3
1. Science—Study and teaching—Evaluation. 2. Science—Ability testing. I. Coffey, Janet. II. Douglas, Rowena. III. Stearns, Carole.
LB1585.A777 2008
507.1'073—dc22

2008018485

Contents

Section 3

Section 4

Foreword

Elizabeth Stage
Director, Lawrence Hall of Science
University of California, Berkeley

It is all too common to pick up a newspaper and see an article about student achievement (usually declining test scores) or district testing policies and the effects of No Child Left Behind on the allocation of instructional time. All around the country, annual testing for the purpose of accountability is dominating public conversations about education. This focus on accountability testing is just one of many assessment responsibilities teachers juggle daily, and probably the least important for supporting student learning. As the essays in this book attest, teachers also need to assess students to guide daily instructional decisions, to promote their further learning, and to assign grades. In a more perfect world, assessment for accountability and assessment for student learning would align, reinforcing one another. Unfortunately, more often than not, such synergy remains elusive.

In 2005, NSTA invited a distinguished group of researchers and teacher educators to share their current research and perspectives on assessment with an audience of teachers. As the conference demonstrated, a rich body of research on what works and what does not is available to inform teachers' assessment practices. The conference also demonstrated the value of an open dialogue among researchers and teachers on practical applications of assessment research to practice. The goal of this book, with chapters by the conference presenters, is to share these research-based insights with a larger audience and to help teachers bring together different assessment priorities and purposes in ways that ultimately support student learning.

This book is also a call for greater teacher involvement in assessment discussions, particularly at the state and local levels. Just as we know from classroom-based research that teachers can gain great insight by listening carefully to their students, so too researchers and policy makers will be better informed by listening to teachers—to the questions they have, the

realities they face, and the dilemmas with which they struggle. Teachers should actively engage in conversations, participate in test design and item development, and help improve the assessment literacy of students and parents. Indeed, teachers' voices are prominent in many of the research efforts described in this book; teachers co-authored many of the chapters. Insights from teachers will help generate strands of research that contribute to richer understandings of assessment practice and its ultimate influence on student learning. While no simple fixes exist for the seemingly divergent assessment purposes, by working together, teachers and researchers can design powerful assessment contexts that help all students reach deep levels of conceptual understanding and high levels of science achievement.

Introduction

Janet Coffey and Carole Stearns

In an era of accountability, talk of assessment often conjures up images of large-scale testing. Although it dominates attention, annual testing is only a small corner of what occurs in the classroom in the name of assessment. Assessment is the chapter test, the weekly quiz, the checking of nightly homework assignments. Assessment can be the observations made as students engage in an activity or the sense-making of student talk as they offer explanations. It is the teacher feedback offered on the lab report, provided to students as they complete an investigation or after they have completed a journal entry. As all of these things and more, assessment is a central dimension of teaching practice.

As the multiple images of assessment suggest, within any classroom, assessment takes on many forms, and must serve multiple purposes. These purposes include accountability and grading. Another important purpose that has received increasing attention is assessment that supports student learning, rather than solely documenting achievement. Different ways to talk about assessment have emerged. We can talk about its *purpose*, as we just did above. We can talk about the *form* assessment takes—the multiple-choice test, the portfolio, the alternative assessment, the written comments or oral feedback, or the piece of student work. Different *uses* of information gleaned from assessment have led us to talk about assessment *of* learning and assessment *for* learning, or, in assessment terminology, summative and formative assessment. All of these purposes, forms, and functions are important; all are at play in the classroom.

Over the past decade, the National Science Foundation (NSF) has funded numerous research efforts that seek to better understand assessment in science and math education at all levels; the various strategies and systems; and the variety of forms, roles, and contexts for assessment *of* and *for* student learning. NSF has also funded assessment-centered teacher professional development efforts and creation of models for assessment systems that seek synergy among different purposes. In 2005–2006, the National Science Teachers Association convened two full-day conferences to help

disseminate these NSF-funded research findings to practitioners. Many of the recipients of those grants shared their work at the conferences and have prepared chapters for this book in an effort to build connections between research and practice and to facilitate meaningful conversation.

Conversations between research and practice are not commonplace, yet greater exchange is essential. Practitioners help researchers better understand the terrain, including the practitioners' underlying rationales for their everyday decision making. These insights from those "on the ground" can inform research and contribute to generative lines of questioning. Although starting points and perspectives may differ, ultimately the assessment research and practitioner communities are working toward the same goal: to better understand the relationships between assessment and learning in order to create classroom environments that support our students' learning.

Researchers are afforded the luxury of stepping back; they can extract a part from the whole—the formative from the summative, for example. They can focus on particular strategies or activities, such as use of notebooks or assessment of lab reports. Teachers, on the other hand, need to make sense of assessment in all its complexity and juggle what may seem like competing priorities and purposes. There may even be times when the different roles teachers take on with respect to assessment appear to conflict: They are, at once, judge and juror, coach and referee. Teachers are asked to figure out ways to navigate these different roles and to align strategies to priorities. They are asked to implement assessment activities and strategies in such a way that a variety of purposes is served, and served well, while mitigating tensions that appear unavoidable.

Research does not hold all the answers. The research community still wrestles with very real and difficult issues that teachers face every day, such as equitable assessments, challenges associated with wide-scale professional development, and assessment designs that capture the complexity of disciplinary reasoning and understanding. As the education community makes progress on these fronts, new challenges and questions arise. No silver bullet exists, nor does a one-size-fits-all fix. However, research can offer insights into strategies and features that are particularly productive, and into frameworks that are particularly compelling.

The essays in this collection will introduce readers to some of the many voices in the national discourse on science assessment, a field currently at the crossroads of education and politics. The essays present authors' deeply held values and perspectives about the roles of assessment and how assessment must not only provide accountability data but also support the learn-

ing of students from different backgrounds. Readers will notice that many of the research studies are grounded in classroom practices and involve teachers as collaborators or in professional development settings. Practitioners' expertise in understanding the complexity of classrooms is crucial to realizing the importance of assessment in deep science learning for all students.

You will not hear a message of consensus here. The research community does not speak in a unitary voice—beyond the claims that there exists a tight coupling between assessment and student learning and that events and interactions that occur in classrooms in the name of assessment do matter. This is not a "how-to" manual. You will not find polished strategies or assessments to try tomorrow in your classroom. Research cannot offer assistance in that form. Strategies, approaches, and frameworks will need modification and accommodation in order to be meaningfully integrated into your classroom and school. As you read, we encourage you to reflect on your own practice, consider your own priorities, and make sense of what you are learning in light of your own school community

Organization of the Book

The chapters in this book are grouped into four sections: (1) formative assessment in the service of learning and teaching; (2) classroom-based strategies for assessing students' science understanding; (3) high-stakes tests; and (4) assessment-focused professional development.

Each section begins with a brief introduction and overview of the included chapters. The section introductions also offer a set of framing questions intended to help readers identify important themes and construct take-home messages that are relevant to their own teaching environment and needs.

The first section, "Formative Assessment: Assessment for Learning," introduces three perspectives on formative assessment: its role in improving student learning; research examining connections between a sequence of formative assessments and their impact on teaching and learning; and the importance of probing how students learn and their misconceptions. Many of the book's central ideas are introduced in this section:

- Roles of assessment in teaching and learning,
- Characteristics of meaningful assessment items,
- Need for research to validate assessment practices,
- Significance of assessing both the knowledge and misconceptions of students,

- Value of assessing students' ability to apply their knowledge, and
- Importance of assessment-focused professional development.

The opening chapter defines classroom-based formative assessment as an ongoing activity informing daily instructional decisions and accompanied by meaningful feedback to students. The author asserts that an essential precursor to raising student achievement in science is providing professional development that will help teachers improve their assessment practices, a topic addressed in many of the chapters and explored in great detail in Section 4.

A research study on correlations between use of embedded formative assessments, teacher practice, and student achievement is the subject of Chapter 2. The focus of the third chapter is the importance of knowing *how* students learn and the nature of their misconceptions. Readers will learn about tools the authors developed to gather and analyze this information.

The chapters in Section 2, "Probing Students' Understanding Through Classroom-Based Assessment," present specific classroom-based strategies for assessing students' science knowledge and understanding. Several of these strategies are closely linked with students' literacy and communication skills, primarily writing, but also drawing, reading, and oral communication. These chapters address the day-to-day issues that teachers confront, such as "How much do my students understand?" "What still confuses them?" "How can I encourage them to communicate more clearly?" and "What constitutes a good formative assessment?"

Several authors write about using familiar classroom artifacts such as students' drawings and notebook entries for assessment purposes. There is a chapter on teaching students to construct reasoned scientific explanations based on their own observations and analysis of data. Secondary teachers may be particularly interested in the chapter on assessing laboratory work. One chapter reports a research study on the use of science notebooks to assess English language learners. (Chapters in later sections also address the needs of English language learners, one in the context of eliminating bias in test items [Chapter 12] and another in a large-scale study of correlations between the science achievement of non-native speakers and the amount of assessment-based professional development their teachers receive [Chapter 17].)

Many of the chapters in this section consider assessments based on familiar classroom routines and artifacts (e.g., science notebooks, lab reports, conversations with and among students) that, when observed through an assessment lens, reveal valuable information about what and how students

are learning. Other chapters in this section describe classroom-based assessment formats and items that were developed by researchers and subjected to field testing in multiple classroom settings. A team of developers describes a suite of formative assessment tasks designed to monitor student learning at several points during a multi-week unit of study. Another chapter introduces a technology-based assessment system developed to track students' problem-solving skills as they interact with a computer simulation. This section concludes with a chapter offering teachers guidelines on constructing standards-based formative assessment probes.

Section 3, "High-Stakes Assessment: Test Items and Formats," begins with an examination of the cognitive demands of several high-stakes test item formats. Authors focus on what students must know and be able to do to succeed on high-stakes tests and how teachers' own classroom assessment can help students meet these challenges. The opening chapter takes readers through the process of designing and field testing items that are closely linked to specific standards-based learning goals. The next chapter analyzes constructed-response test items, a format commonly used in national and international tests, such as TIMSS and NAEP. The authors present sample items and detailed scoring guides to help teachers better understand how such items are scored. Another chapter provides teachers with guidelines for analyzing the content and format of high-stakes test items and creating closely aligned questions to use in their own classrooms.

Section 3 continues with a chapter summarizing the National Research Council's (NRC) report on design principles for state-level science assessment systems. The authors discuss the goals of state-level assessment, calling attention to the distinct differences between these tests and the classroom-based assessments described in Section 2. The concluding chapter offers reflections by a literacy expert on high-stakes testing practices and test items in his field. He summarizes the lessons learned and offers some suggestions to science test developers.

In Section 4, "Professional Development: Helping Teachers Link Assessment, Teaching, and Learning," authors describe several large-scale professional development initiatives that emphasize building assessment expertise. Programs in Seattle, Washington, Toledo, Ohio, Miami, Florida, and Colorado Springs, Colorado are discussed. While each had a different approach to professional development design, all included a research component investigating potential correlations between the teachers' experiences and their student performance on high-stakes tests. Each study reports compelling data showing a positive correlation between teachers' participation in the professional development efforts and student achievement on high-stakes science tests.

A chapter on a classroom observation research tool titled the Reform Teacher Observation Protocol (RTOP) offers another approach to professional development. The authors discuss the use of this tool by secondary teachers to self-evaluate their classroom assessment practices. The final chapter describes strategies that school teams can use to analyze assessment data from multiple sources; including high-stakes tests, classroom-based assessments, and teacher observations, for the purposes of program evaluation and guiding instructional decisions.

* * *

This brief summary does little justice to the richness of the essays herein and to the multiple examples of meaningful science assessment practices they explore. The collection reflects work with socioeconomically and ethnically diverse populations to better understand the attributes of equitable assessment practices. While the authors may describe an assessment study conducted within a narrow context (science teachers will recognize the constraints required by a controlled experiment), the findings and recommendations are broadly applicable. For example, professional development programs in Seattle, Washington, offer many ideas equally relevant for schools and districts in other parts of the United States. Similarly the assessment potential of student notebooks extends far beyond classrooms in El Centro, California.

We hope that this book can be used to fuel the conversations about assessment sparked in the initial NSTA conference. From the informal interactions that occur among students and teachers to more formal exchanges, from item design to grading, and from classroom systems of reporting on progress to large-scale external state tests, fodder exists for deep and provocative discussion. In the essays that follow, readers have an opportunity to consider the issues closely and to reflect on the ways in which assessment can be better coordinated. We hope that, eventually, the entire system will become more synergistic in order to meet the many purposes of assessment while not neglecting or undermining any single one.

The editors are grateful to the researchers who contributed to this volume for their commitment to communicating their work to practitioners, the ultimate consumers of science assessment knowledge. We hope that readers will find many ideas that enrich their own understanding of the assessment landscape and help them better serve their students. We encourage teachers to actively engage in the national assessment conversation and to share insights they develop in their own classrooms.

SECTION 1

Formative Assessment: Assessment for Learning

Section 1 focuses on the nature of formative assessment and its multiple roles in the classroom: monitoring student learning, guiding instruction, providing meaningful feedback to students, engaging them in self assessment, and revealing their misconceptions.

Discussion Questions

- What does formative assessment look like in the science classroom?
- In what ways can formative assessment impact student learning?
- What is the relationship between formative assessments and high-stakes tests?
- What can formative assessments teach us about how students learn and how to help them become better learners?

Chapter Summaries

The essay by Dylan Wiliam, Institute of Education, University of London, is based on his extensive experience with classroom teachers. Wiliam asserts that an essential step to raising student achievement in science is to provide professional development that will help teachers improve their formative assessment practices. He envisions classroom assessment as an ongoing activity informing daily instructional decisions and accompanied by meaningful feedback to students. Wiliam emphasizes his thesis with a

sports analogy. Coaches observe their athletes and then provide both diagnostic and formative feedback to improve their performance. The diagnostic helps the athlete identify what is wrong and the formative is a guide to making it better.

Richard Shavelson and his colleagues at Stanford University and the Curriculum Research and Development Group at the University of Hawaii report on their recent study with middle school physical science teachers. After developing a series of embedded formative assessments closely aligned with the local curriculum, they conducted controlled experiments investigating the impact of the assessments and the information these assessments provided on students' knowledge and motivation to learn science. The chapter offers readers a candid view of a multifaceted research project, the complexity of controlling the research variables in a real-world situation, the findings (many of them unexpected), and the challenges associated with classroom-based research.

Jim Minstrell recounts his transition from science teacher to researcher. He describes his early career as a high school physics teacher and his gradually turning his attention from *what* his students were learning to *how* they were learning. Now a teacher-researcher, Minstrell talks about the Diagnoser, a tool he and his FACET Innovations colleagues—Ruth Anderson, Pamela Kraus, and James E. Minstrell—developed to explore the nature of students' commonly held misconceptions and to extend his research on how students develop scientific understanding.

CHAPTER 1

Improving Learning in Science With Formative Assessment

Dylan Wiliam
Institute of Education, University of London

In recent years, the No Child Left Behind law has focused attention on student achievement in science across the United States, but there are more important reasons for being concerned with student achievement. Increasing student achievement brings substantial benefits both to the individual and to society. For the individual, improved school achievement increases career earnings, improves health and well-being, and actually lengthens life (Wiliam and Thompson 2007; Lleras-Muney 2005). For society, the benefits include increased tax revenues, savings in public health costs, reduced law-enforcement and prison costs, as well as savings in welfare budgets (Hoff 2007).

Achievement in science in particular is likely to be increasingly important in the future for the needs of employment, but it will also be essential for democratic citizenship. Without an understanding of what science can do (and what it can't) and how science does what it does, public policies about issues such as genetically modified foods, assisted reproductive technologies, and human cloning are likely to be set on the basis of populist journalism rather than scientific evidence.

The focus of this chapter, therefore, is about how we can raise achievement in science in the United States. In this chapter I will argue that if we are serious about raising achievement in science, then we need to look beyond "what works" in education to notions of cost-benefit—not just whether a particular initiative raises achievement, but by how much, and at what cost. I will show that the evidence currently at hand suggests that

this is done by investing in teacher professional development, but of a sort very different from what occurs in most school districts. I will also show that this professional development needs to be focused specifically on changing what teachers do in the classroom and, in particular, needs to be aimed at changing teachers' minute-by-minute and day-by-day use of assessment to modify instruction, sometimes called formative assessment or assessment for learning.

Value for Money in Education Reform

Many policy makers have focused on "what works" in education, but such a focus is misguided, since what is most important is the size of effect relative to cost. An intervention might "work" but the effects might be too small to be worth bothering with, or it might produce substantial effects but be too expensive to implement. The focus in school improvement needs to be on the ratio of the size of the benefit to the cost incurred in bringing it about.

When we adopt this perspective, we find that some effective interventions are too costly, and some interventions that have only a small impact on student learning nevertheless turn out to be implementable at a modest cost. To take one example, the research shows clearly that reducing class size raises student achievement. So what? The important questions should be how much improvement, and at what cost? And here the data are depressing.

Jepsen and Rivkin (2002) found that reducing elementary school class size by 10 students would increase the proportion of students passing typical mathematics and reading tests by 4% and 3% respectively. While there is evidence that the effects are larger for the early years, beyond this point, class-size reduction appears to be a very expensive way of increasing student achievement. To make this more precise, consider the cost of this intervention per classroom of 30 students. Reducing class-size from (say) 30 to 20 would increase the salary costs by approximately 50%, because we would need three teachers instead of two for a group of 60 students. Assuming an average teacher salary cost of $60,000 per year, this would cost $30,000 per year to implement.

Using tests such as those used in international comparisons such as TIMSS (Rodriguez 2004) as a benchmark, we find that the effect of these class-size reductions would be equivalent to students in the smaller classes learning in three months what it would take the students in the larger classes four months to learn; a 33% increase in the speed of learning.

A more recent study by Jenkins, Levacic, and Vignoles (2006) in England found that additional resources, if used to reduce class size, might have a larger increase on student achievement in middle and high school science (up to four times the effect size found by Jepsen and Rivkin), but this model assumed that additional science teachers of equivalent quality to those already on the job could be found, which is at the very least questionable (one of the key factors in the modest results in the Jepsen and Rivkin research cited above was that the extra teachers were not as good as those already in place). Nevertheless, it may be that there are special cases, such as early years, and secondary science, where class-size reduction may have a substantial effect on student achievement. However, in general, class-size reduction would appear to offer only modest increases in student achievement at very high cost.

How else could we obtain the same increase in student achievement? One obvious candidate is to increase teacher subject knowledge. Hill, Rowan, and Ball (2005) found that an increase of two standard deviations in what they termed "mathematical knowledge for teaching" was associated with an increase of up to 0.1 standard deviations in student achievement in mathematics. This was not a direct experiment, and so we cannot infer a causal relationship. However, it suggests that the same improvement in achievement that would be gained by reducing class size from 30 to 20 might be secured by increasing teacher subject-matter knowledge by two standard deviations. Unfortunately, there is currently little evidence about how to do this.

In contrast, supporting teachers in developing the use of assessment for learning has been shown to roughly double the speed of learning (Wiliam et al. 2004; Hayes 2003). In other words, students learned in six months what would have taken a year to learn in other classrooms. The cost of this intervention was around $8,000 per teacher, but of course, unlike class-size reduction, the cost of which has to be found annually, investing in teachers' professional development is a one-off expenditure, which can be depreciated over the teacher's remaining career. Against this, some cost of annual renewal needs to be allowed. Assuming that a new teacher will continue to teach for at least five years, and allowing time for four hours of meetings with colleagues per month for renewal, the cost of teacher professional development focused on assessment for learning would appear to be around $3,000 per teacher (and therefore per classroom) per year.

Compared to class-size reduction, therefore, improving teachers' use of assessment for learning would appear to promise two or three times the increase in student learning, for around one-tenth the cost. Even in the special case of secondary school science discussed above, assessment for learning produces roughly the same size of benefit as reducing class size by 30%, at less than one-fifth the cost.

Assessment for Learning

When teachers are asked how they assess their students, they typically talk about tests, examinations, quizzes, and other formal methods. When they are asked how they know whether their students have learned what they have taught, the answers are very different. They talk about homework, classwork, the things students say in classroom discussions, and even the expressions on their faces. This is the great disconnect in education worldwide. Assessment that serves the needs of teachers and their students is seen as completely separate from, and indeed, incompatible with, assessment that serves the needs of parents, administrators, policy makers, and other stakeholders.

At one extreme we have a teacher questioning a student, trying to elicit evidence of (mis)conceptions that are likely to impede future learning. At the other extreme we have the use of Advanced Placement tests, used both to give students college-level credit and, by some universities, to decide which students to admit. The obvious conclusion is that the latter kind of assessment hinders learning while the former helps, but things are not that simple.

The teacher's questioning of the student can cause damage, possibly irreparable, to the student's sense of self if undertaken in a humiliating way. And at the other extreme, Advanced Placement tests provide clear guidance to teachers and students about what the vague words in examination syllabuses mean. Furthermore, when used as "trial" examinations, sample papers allow students to benchmark themselves against the standard established by the College Board and to help each other rectify deficiencies. If we are to design assessment systems that help rather than hinder learning, we must go beyond looking at the assessments themselves and look at deeper issues about how the assessments help learners and their teachers know where the learners are in their learning, where they are going, and how to get there.

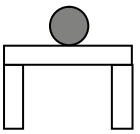
Through extensive reviews of the available research evidence, and through extensive fieldwork with teachers, both in the UK and in the United States (see the suggestions for further reading at the end of this chapter), we have identified five "key strategies," which, when implemented appropriately, allow assessment to help, rather than hinder, learning.

1. Engineer effective classroom discussions, questions, activities, and tasks that elicit evidence of student learning.

The first step in using assessment to help learning is to collect the right sort of evidence, and here it is clear that the tools that teachers use to find out where students are in their learning are given too little attention.

Few teachers plan the kinds of tasks, activities, and questions that they use with their students specifically to elicit the right kind of evidence of student learning. As an example of a good question, consider the science question shown in Figure 1.1.

Figure 1.1 Science Item



The ball resting on the table is not moving. It is not moving because
A. no forces are pushing or pulling on the ball.
B. gravity is pulling down, but the table is in the way.
C. the table pushes up with the same force that gravity pulls down.
D. gravity is holding it onto the table.
E. there is a force inside the ball keeping it from rolling off the table.

Source: Adapted by the author from Wilson, M., and K. Draney. 2004. Some links between large-scale and classroom assessments: The case of the BEAR assessment system. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education.* Chicago, IL: University of Chicago Press.

Response A clearly relates to the well-known misconception that if there is no movement, then there are no forces acting. Response B is more complex. In one sense there is nothing wrong with B. After all, gravity *is* pulling the ball down, and the table *is* in the way. The reason that B is a less preferable response to C is not that it is incorrect; it is that it is not physics. Science provides us with a powerful, but unnatural, way of thinking about the world; after all, if it were natural, we would not need to teach it. Students who choose B rather than C have not understood the important idea that lack of movement denotes forces in equilibrium, not the absence of forces.

By crafting questions that explicitly build in the under- and overgeneralizations that we know students make, we can get far more useful information about what to do next. By equipping each student in the class with a set of cards with A, B, C, D, and E on them, and by requiring all students to respond simultaneously with their answers, the teacher can generate a very solid evidence base for deciding whether the class is ready to move on. If every student responds with C, then the teacher can move on with confidence that the students have understood. If everyone simply responds with A, then the teacher may choose to re-teach some part of the topic. The most likely response, however, is for some students to respond correctly and for others to respond incorrectly. This provides the teacher with an opportunity to conduct a classroom discussion in which students with different views can be asked to justify their selections. Moreover, because the teacher knows which student gave which response, the discussion can be better orchestrated (e.g., "Shane, you also chose B. Did you choose it for the same reason that Alicia gave, or a different reason?").

Of course, planning such questions takes time, but by investing the time before the lesson, teachers are able to address students' confusion during the lesson, with the students still in front of them. Teachers who do not plan such questions are forced to put children's thinking back on track by giving extended comments while grading, thus dealing with the students one at a time, after they have left the classroom.

2. Provide feedback that moves learning forward.

The research on feedback shows that much of the feedback that students receive has, at best, no impact on learning, and can actually be counterproductive. One extraordinary study (Kluger and DeNisi 1996) reviewed

more than 3,000 research reports on the effects of feedback in schools, colleges, and workplaces. They then rejected studies that failed to reach the highest standards of methodological rigor and were left with just 131 studies. Across these 131 studies, they found that, on average, feedback did increase achievement, but that in 50 of the studies (i.e., almost two in five), feedback actually made people's performance worse than it would have been without feedback. The key feature of these studies was that feedback was, in the psychological jargon, "ego-involving." In other words, the feedback focused attention on the person rather than the quality of the work—for example, by giving scores, grades, or other forms of report that encouraged comparison with others. For the 81 studies that found a positive impact on performance, Kluger and DeNisi found that the biggest impacts occurred when feedback told not just what to do to improve, but also how to go about it.

An example from athletics may be helpful here. If a young fast-pitch softballer has an ERA (earned run average) of 10, we know that she is not doing well. This is the *monitoring* assessment. The monitoring assessment identifies that there is a problem, but doesn't identify what it is. By looking at her pitching, we might see that the reason that her ERA is so high is that she is trying, unsuccessfully, to pitch a rising fastball. This is a pitch that is thrown with enough backspin so that as it reaches the batter, it rises sharply, making it almost unhittable. The problem is that if it does not rise, then the result is a fastball over the center of the plate, which is very easy to hit, and this is exactly what is happening to our pitcher. Her rising fastball is not rising. This is the *diagnostic* assessment. The diagnostic assessment identifies where the problem is, but by itself, doesn't give the athlete any clue about how to go about making improvements. However, if the pitching coach can see that the reason that the pitcher is struggling to pitch the rising fastball is because she is not dropping the pitching shoulder low enough to deliver the pitch from below the knee, then this gives the athlete something to work with. This is the *formative* assessment. Just as we use the term *formative* to describe the experiences that shape us as we grow up, a formative assessment is one that shapes learning. Much of the feedback that students get while learning science is no more helpful than telling our fast-pitch softballer to make sure her rising fastball rises or telling a bad comedian that he needs to be funnier. If feedback is to impact learning it must focus on what needs to happen next; in other words, it must be a *guide*

to action and not just a demand for it. For an example of a feedback system focused on middle school science, see Clymer and Wiliam (2006/2007).

3. Clarify and share learning intentions and success criteria with learners.

In an article titled "The View From the Student's Desk," written more than 35 years ago, Mary Alice White (1971) said:

> *The analogy that might make the student's view more comprehensible to adults is to imagine oneself on a ship sailing across an unknown sea, to an unknown destination. An adult would be desperate to know where he [sic] is going. But a child only knows he is going to school.... The chart is neither available nor understandable to him.... Very quickly, the daily life on board ship becomes all important.... The daily chores, the demands, the inspections, become the reality, not the voyage, nor the destination.* (p. 340)

In a similar vein, I have walked into many science classrooms and asked students what they were doing, only to be told something like "page 34,"as if that were all I needed to know. For many students, school is just a series of tasks whose purposes are unclear and even what counts as success is mysterious, especially for students from less-advantaged backgrounds.

One study that is particularly notable for reducing the achievement gap between lower- and higher-achieving students was conducted by White and Frederiksen (1998). The study involved three teachers, each of whom taught four parallel seventh-grade classes in two U.S. schools. The average size of the classes was 31 students. In order to assess the representativeness of the sample, all the students in the study were given the Iowa Test of Basic Skills (ITBS), and their scores were close to the national average. All 12 classes followed a novel curriculum (called ThinkerTools) for 14 weeks. The curriculum had been designed to promote thinking in the science classroom through a focus on a series of seven scientific investigations (approximately 2 weeks each).

Each investigation incorporated a series of evaluation activities. In two of each teacher's four classes, these evaluation episodes took the form of a discussion about what they liked and disliked about the topic. For the other two classes, they engaged in a process of "reflective assessment." Through a series of small-group and individual activities, the students were introduced

to the nine assessment criteria (each of which was assessed on a 5-point scale) that the teacher would use in evaluating their work. At the end of each episode within an investigation, the students were asked to assess their performance against two of the criteria. At the end of the investigation, students had to assess their performance against all nine. Whenever they assessed themselves, they had to write a brief statement showing which aspects of their work formed the basis for their rating. At the end of each investigation, students presented their work to the class, and the students used the criteria to give one another feedback.

When the researchers analyzed the achievement of the students, the weakest students in the "reflective assessment" group performed as well as the best students in the control group, and the other students did even better, the result of which was to reduce by half the achievement gaps in the "reflective assessment" classes.

4. Activate students as owners of their own learning.

One of the great traps of teaching is the belief that teachers create learning. This is particularly important when teachers are under pressure to improve student results, because studies have shown that when teachers are told they are responsible for making sure that their students do well, the quality of their teaching deteriorates, as does their students' learning (Deci et al. 1982); hence the old joke about schools being places where children go to watch teachers work.

Only learners create learning, and so, when we look at the role that assessment plays in promoting learning, the crucial feature is not the validity of the assessment, or its reliability, but its impact on the student. No matter how reliable or valid the assessment is, if it communicates to students that they cannot learn, it will hinder learning. Particularly important here is the work of Carol Dweck, who over a 30-year period has examined the way that students make sense of their successes and failures in school (see Dweck 2000, for a very readable summary of this huge volume of work). As a result of their experiences, some students come to believe that ability is fixed. The reason that this is so injurious to future learning is that every time students with this belief are faced with a challenging task, their first reaction is to engage in a calculation about whether they are likely to succeed or not. If they feel confident that they will succeed, or if they feel confident that the task is so hard that many others will also fail, they will

attempt the task. However, if they feel that there is a danger that they will fail while others will succeed, they will disengage in order to protect their sense of self. Put simply, they are deciding that they would rather be thought lazy than dumb. It is the same choice that most adults would make.

There are other students, who, for a variety of reasons, have come to regard ability as incremental rather than fixed. They believe that "smart" is not something you *are* but something you *get*. For these students, challenging tasks are opportunities to increase their abilities, so whether their beliefs in their chances of success are high or low, they engage with a task in order to grow. What is particularly interesting is that the same student can believe the ability in science is fixed, while seeing ability in sports or music as incremental. Most students believe that ability in, for example, triple jump, throwing the javelin, or guitar playing can be improved by practice. We need to inculcate the same beliefs about science.

In general, we need to activate students as owners of their own learning, so that they see challenge as a spur to personal growth, rather than as a threat to self-image. We need students who own their learning to the extent that they can self-manage both their emotional and their cognitive responses to challenge, so that all their energies are spent on developing capability rather than disguising its absence (see Wiliam 2007 for a summary of research in this area).

5. Activate students as learning resources for one another.

The research on collaborative learning is one of the success stories of education research. Research in many areas of education produces ambiguous or contradictory findings whereas the research on collaborative learning, most notably the work of Robert Slavin (Slavin, Hurley, and Chamberlain 2003), has shown that activating students as learning resources for one another produces some of the largest gains seen in any educational interventions, provided two conditions are met. The first is that the learning environment must provide for group goals, so that students are working *as* a group, rather than just working *in* a group. The second is individual accountability, so that each student is responsible for his or her contribution to the whole, so there can be no "passengers."

With regard to assessment, then, a crucial feature is that the assessment encourages collaboration among students while they are learning. To achieve this, the learning intentions and success criteria must be accessible

to the students (see above), and the teacher must support the students as they learn how to help each other improve their work. One particularly successful format for doing this has been the idea of "two stars and a wish." The idea is that when students are commenting on each others' work, they do not give evaluative feedback, but instead have to identify two positive features of the work (two "stars") and one feature that they feel merits further attention (the "wish").

Teachers who have used this technique with students as young as five years old have been astonished to see how appropriate the comments are, and, because the feedback comes from a peer rather than someone in authority over them, the recipient of the feedback appears to be more able to accept the feedback (in other words, students focus on growth rather than preserving well-being). In fact, teachers have told us that the feedback that students give each other, while accurate, is far more hard-hitting and direct than they would themselves feel able to provide. Furthermore, the research shows that the person providing the feedback benefits just as much as the recipient, because he or she is forced to internalize the learning intentions and success criteria in the context of someone else's work, which is less emotionally charged than doing so in the context of one's own work.

The "Big Idea": Keeping Learning on Track

The "big idea" that ties these strategies together is that assessment should be used to provide information to be used by students and teachers that is then used to modify instruction in real time in order to better meet student needs. In other words, assessment is used to keep learning on track.

That this is not common practice can be seen by imagining what would happen if an airline pilot navigated the way that most teachers teach. The pilot would set a course from the starting point (say New York) to the destination (say San Francisco). The pilot would then fly on this heading for the calculated time of travel, and then, when that time had elapsed, would land the plane at the nearest airport, and upon landing ask, "Is this San Francisco?" Worse, even if the plane had actually landed in Sacramento, the pilot would require all the passengers to leave, because he had to get on to his next job.

This would be absurd, and yet, this is how many teachers teach. They teach a unit for two or three weeks, and at the end of that teaching, they assess their students. And whatever the result of that assessment, the teacher

is then on to the next unit, because of the district's pacing guide. If we are to keep learning on track, assessment cannot wait until the end of the unit. Instead, like the pilot, the teacher plans a course but then takes frequent readings along the way, adjusting the course as conditions dictate.

Substantial increases in student achievement are possible if we can increase the amount of assessment for learning in classrooms, but achieving this is no easy task.

Putting It Into Practice: The Case for Teacher Learning Communities

The fact that we know what needs to be done does not, of course, mean that we know how to do it. While the work of Black et al. (2003) has shown what we can achieve, the track record of professional development in producing significant effects on a large scale is rather unimpressive. However, this should not worry us unduly because very little of the professional development that teachers have received in the past is consistent with what we know makes for effective teacher change (Wiliam and Thompson 2007).

Why this is the case is complex, and beyond the scope of this article (see Wiliam 2003 for an extended discussion). What is clear is that in general, researchers have underestimated the complexity of what it is that teachers do; in particular, researchers have failed to understand how great an impact context has on teachers' practices. That is why "What works?" is not the right question, because everything works somewhere, and nothing works everywhere. The right question is "Under what conditions will this work?" And even if we might be able to answer this question scientifically at some point in the future, we are so far away from an answer now that we have to rely on the professional judgment of teachers.

This is why we cannot tell teachers what to do. This conclusion does not stem from a desire to be nice to teachers. Indeed, if I could identify a way of telling teachers what to do that would raise student achievement, I would have no hesitation in mandating it if I had the power to do so. Schools exist for the students, not to provide employment opportunities for teachers. The reason that we cannot tell teachers what to do is that we cannot provide them with reliable guides to action. The situations that they face in their classrooms are just too varied for us to predict. That is why professional judgment is important; we have to develop the ability of

teachers to react appropriately to situations for which they have not been specifically prepared.

The specific changes that I am arguing for here appear to be quite difficult for teachers to implement because they involve changing habits. A teacher with 20 years experience may well have asked something like half a million questions in her career. And when you've done something the same way half a million times, it's quite difficult to start doing it in any other way. But there is a deeper reason why change is difficult, even for inexperienced teachers. Teachers learn most of what they know about teaching before their 18th birthday. In the same way that most of us learn what we know about parenting through being parented, teachers internalize the "scripts" of school as students. Even the best four-year teacher education programs will find it hard to overcome the models of practice that student teachers will have learned in the 13 or 14 years that they spent in school.

This is why, if we are to have any chance of really changing teacher practice, we have to take seriously how hard this is going to be. We are asking teachers to change the routines and the practices that get them through the day; in the transition, they will get worse before they get better. Indeed, many of the teachers we have worked with described making these changes as "scary." They saw involving the students more in their learning as requiring them to "give up control" of their classrooms. A year later, the same teachers described the process as one of sharing responsibility for learning with the learners—the same process, but viewed in a radically different perspective.

The process that I believe provides the best mechanism for supporting teachers in making these changes is through the use of teacher learning communities (TLCs). I say this not out of some ideological commitment to the benefits of teachers talking to each other, but because of the nature of the changes we are seeking to produce. If we were trying to increase teacher subject knowledge, then TLCs would not be a very sensible approach—it would be far better to arrange for high-quality direct instruction. But when we are trying to change deeply ingrained, routinized practices or habits, then it seems that TLCs offer the best hope, and indeed, the results we have achieved in the United States have been very encouraging.

Over the past three years, we have tried a number of different approaches to establishing and sustaining TLCs, and as a result of this experimentation, it appears to us that five principles appear to be particularly important: gradualism, flexibility, choice, accountability, and support.

Gradualism

Asking teachers to make wholesale changes in their practices is a little like asking a golfer to change her swing during a tournament. Teachers have to maintain the fluency of their classroom routines, while at the same time disrupting them. Teachers should develop an action plan that specifies a small number of changes—ideally two or three—that they will make in their teaching. As teachers establish new techniques in their practices, they can take on additional ones. For administrators, there will be a temptation to push teachers to change faster than they might otherwise do, but the result will be only a shallow adoption of the new practices as long as the teachers are being monitored. As soon as the supervision is relaxed, the teachers will revert to their earlier practices, and nothing will have been achieved. Even districts in a hurry will have to hasten slowly.

Flexibility

Teachers will need to modify techniques to make them work in their classrooms; in the process of adapting techniques, teachers often refine and improve them. One high school mathematics teacher heard about the "traffic light" technique in which at the end of a piece of work, students indicate their confidence in their understanding of a piece of work with a green, yellow, or red circle, representing complete, partial, or little understanding. She decided that she would not wait until the end of the lesson to engage students in this self-assessment and gave each student a disk, green on one side and red on the other. At the start of each lesson students place the disk on their desk with the green side showing. A student can indicate confusion at any time by turning the disk over to show red. The teacher found that students who had never asked a question all year in class were prepared to signal their confusion in this way.

Another teacher tried this approach, but found it difficult to see the disks from the front of the class so she provided each student with three paper cups—one green, one yellow, and one red—nested inside each other on the students' desks. Students used these cups to indicate whether they were following the teacher's explanation (green), wanted the teacher to slow down (yellow), or wanted her to stop in order to ask a question (red). The teacher made students accountable for signaling correctly by establishing a rule that whenever one student showed red, a student who was showing green or yellow would be chosen at random to come to the front of the class

to answer the question posed by the student showing red. In this classroom there is nowhere to hide!

Choice

As noted above, teachers often describe the process of changing their practices as "scary," but when they are responsible for choosing what they will change about their practices they feel empowered, especially when they can choose from a range of techniques that appeal to them. This choice lies, however, within a framework of accountability. While teachers are free to choose what they change, they are accountable for changing something.

Accountability

Most professionals involved in teacher development will have had the experience of generating considerable enthusiasm for, and commitment to, change during a workshop, only to find that all the good intentions seem to be erased once the teachers return to the classroom. Teachers should be held accountable for making changes by colleagues at monthly meetings of their teacher learning community. Each teacher describes what he or she tried and how it went. Teachers repeatedly tell us that having to face their colleagues helps them move their "change" task to the top of their in-box.

Support

Along with ideas for what to change and the support of a teacher learning community, two elements are highly desirable, if not essential, for teacher learning. The first is training for those who will lead the learning communities. The person leading the learning community must be clear about his or her role. The role of the leader is not to create teacher change but to engineer situations in which the teacher change can take place. Those in supervisory roles often find this more difficult than do teachers, because those teaching every day understand how difficult it is to change practice.

The second element is peer observation. Collaborative planning in the monthly TLC meetings can help teachers focus on what they want to develop in their practices, but teachers need support in carrying out these resolutions. To distinguish these observations clearly from those routinely carried out to manage performance, these observations should be done by peers rather than supervisors. The teacher being observed must set the

agenda for the observation and spell out for the observer what should count as evidence. By defining the observer's role, both in terms of what is to be looked for and what counts as evidence, the observer's own prejudices are minimized, and the difference between this and supervisory observation is emphasized. (For further details on setting up and sustaining learning communities for formative assessment see Wiliam [2007/2008].)

Integrated Assessment

The available research evidence, as well as our experience of working with many school districts in the United States, suggests that the use of teacher learning communities, focused on the use of minute-to-minute and day-to-day assessment to adjust instruction to meet student needs, represents the most powerful single approach to improving student achievement. However, if we are to maximize the impact on student learning, other parts of the system need to be "in sync." In addition to the minute-by-minute and day-by-day assessments that allow teachers to keep learning on track, teachers also need a range of more formal assessment tasks and activities that support valid and reliable conclusions about the extent of student learning. Our experience is that teacher-made assessments often focus on shallow aspects of learning, rather than the "big ideas." Developing high-quality assessments that involve students and motivate them to improve takes time, but there are good examples of how to go about this (Stiggins, Arter, and Chappius 2004). On a longer timescale, quarterly assessments that are paced to the curriculum can provide school leadership teams with valuable information about the progress—or lack of it—that is being made by students, and annual diagnostic analyses of high-stakes tests can provide important insights into the alignment between the teaching and the national curriculum. None of these different kinds of assessment is in conflict with any of the others. Each represents an important part of a complex machine, providing information at the right level of specificity for the decision that needs to be made. Together they form a balanced assessment system that can produce unprecedented increases in science achievement, benefiting both the individual and society as a whole.

References

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

Clymer, J. B., and D. Wiliam. 2006/2007. What's wrong with the way we grade science? *Educational Leadership* 64(4): 36–42.

Deci, E. L., N. H. Speigel, R. M. Ryan, R. Koestner, and M. Kauffman. 1982. The effects of performance standards on teaching styles: The behavior of controlling teachers. *Journal of Educational Psychology* 74: 852–859.

Dweck, C. S. 2000. *Self-theories: Their role in motivation, personality and development*. Philadelphia: Psychology Press.

Hayes, V. P. 2003. *Using pupil self-evaluation within the formative assessment paradigm as a pedagogical tool.* EdD thesis, University of London.

Hill, H. C., B. Rowan, and D. L. Ball. 2005. Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal* 42(2): 317–406.

Hoff, D. J. 2007. Economists tout value of reducing dropouts. *Education Week* 26(Feb. 14): 5, 15.

Jenkins, A., R. Levacic, and A. Vignoles. 2006. *Estimating the relationship between school resources and pupil attainment at GCSE* (Vol. RR727). London, UK: Department for Education and Skills.

Jepsen, C., and S. G. Rivkin. 2002. *What is the tradeoff between smaller classes and teacher quality?* Cambridge, MA: National Bureau of Economic Research.

Kluger, A. N., and A. DeNisi. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin* 119(2): 254–284.

Lleras-Muney, A. 2005. The relationship between education and adult mortality in the United States. *Review of Economic Studies* 72(1): 189–221.

Rodriguez, M. C. 2004. The role of classroom assessment in student performance on TIMSS. *Applied Measurement in Education* 17(1): 1–24.

Slavin, R. E., E. A. Hurley, and A. M. Chamberlain. 2003. Cooperative learning and achievement. In W. M. Reynolds and G. J. Miller (Eds.), *Handbook of psychology, volume 7: Educational psychology* (pp. 177–198). Hoboken, NJ: Wiley.

Stiggins, R. J., J. A. Arter, and S. Chappius. 2004. *Classroom assessment for student learning: Doing it right—using it well*. Portland, OR: Assessment Training Institute.

White, M. A. 1971. The view from the student's desk. In M. L. Silberman (Ed.), *The experience of schooling* (pp. 337–345). New York: Rinehart and Winston.

White, B. Y., and J. R. Frederiksen. 1998. Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction* 16(1): 3–118.

Wiliam, D. 2003. The impact of educational research on mathematics education. In A. Bishop, M. A. Clements, C. Keitel, J. Kilpatrick, and F. K. S. Leung (Eds.), *Second international handbook of mathematics education* (pp. 469–488). Dordrecht, Netherlands: Kluwer Academic Publishers.

Wiliam, D. 2007. Keeping learning on track: Formative assessment and the regulation of learning. In F. K. Lester Jr. (Ed.), *Second handbook of mathematics teaching and learning* (pp. 1053–1098). Greenwich, CT: Information Age Publishing.

Wiliam, D. 2007/2008. Changing classroom practice. *Educational Leadership* 65(4): 36–42.

Wiliam, D., and M. Thompson. 2007. Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning*. Mahwah, NJ: Lawrence Erlbaum.

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

Wilson, M., and K. Draney. 2004. Some links between large-scale and classroom assessments: The case of the BEAR assessment system. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education* (Part II, pp. 132–154). Chicago, IL: University of Chicago Press.

Further Reading

Black, P. J., and D. Wiliam. 1998. Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan* 80(2): 139–148.

Black, P., and C. Harrison. 2002. *Science inside the black box: Assessment for learning in the science classroom*. London, UK: NFER-Nelson.

Black, P., C. Harrison, C. Lee, B. Marshall, and D. Wiliam. 2004. Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan* 86(1): 8–21.

Clymer, J. B., and D. Wiliam. 2006. Improving the way we grade science. *Educational Leadership* 64(4): 36–42.

Leahy, S., C. Lyon, M. Thompson, and D. Wiliam. 2005. Classroom assessment: Minute-by-minute and day-by-day. *Educational Leadership* 63(3): 18–24.

Wiliam, D. 2006. Assessment: Learning communities can use it to engineer a bridge connecting teaching and learning. *Journal of Staff Development* 27(1): 16–20.

Wiliam, D. 2006. Assessment for learning: Why, what and how. *Orbit: OISE/UT's magazine for schools* 36(3): 2–6.

CHAPTER 2

On the Role and Impact of Formative Assessment on Science Inquiry Teaching and Learning

Richard J. Shavelson, Yue Yin, Erin M. Furtak, Maria Araceli Ruiz-Primo, Carlos C. Ayala
Stanford Educational Assessment Laboratory

Donald B. Young, Miki K. Tomita, Paul R. Brandon, Francis M. Pottenger III
Curriculum Research & Development Group

Science education researchers, like science teachers, are committed to finding ways to help students learn science. Like teachers, we researchers start with an informed hunch about something that we think will improve teaching. Then we work with teachers and try out our hunch in real classrooms. If we get positive results, we share them with a wide range of educators. Sometimes we find out that our hunch does not work, and we try to figure out what went wrong so that we can improve it the next time. In other cases, we find that while the idea may have been good, the technique will not work in practice. In those cases, we continue our search for other ways to help improve students' learning of science.

In reviewing the literature on assessment, Paul Black and Dylan Wiliam found strong evidence that embedding assessments in science curricula would lead to improved student learning and motivation (Black and Wiliam 1998; see also Wiliam, Chapter 1 in this book). Based on this finding, our team of teachers, curriculum and assessment developers, and science education researchers developed a series of *formative assessments* to embed in a middle school physical-science unit on sinking and floating. We wanted

to see if this kind of assessment, which helps teachers to determine the status of students' learning while a unit is still in progress, would improve sixth- and seventh-grade students' knowledge and motivation to learn science. If it worked, we knew we might have a large-scale impact on teaching and learning.

In this chapter, we begin by describing what we mean by formative assessment and outline the potential and challenges of trying to implement and study this promising technique for scientific inquiry teaching. We then describe our study on formative assessment in middle schools, including some mistakes and wrong turns, and what we found when we tested our ideas experimentally. We conclude with future challenges in improving science education with formative assessment.

What Is Formative Assessment?

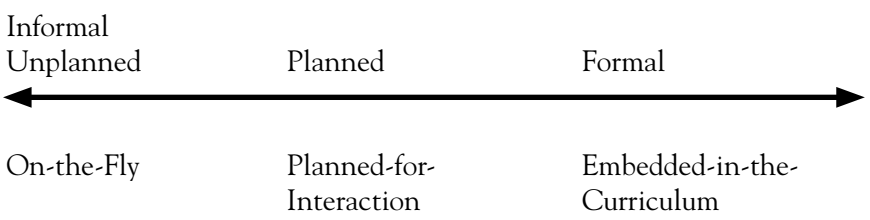
Formative assessment is a process by which teachers gather information about what students know and can do, interpret and compare this information with their goals for what they would like their students to know and be able to do, and take action to close the gap by giving students suggestions as to how to improve their performance. In this way, formative assessment is carried out for the purpose of improving teaching and learning while instruction is still in progress.

To clarify what we mean by *formative assessment*, consider the large-scale, high-stakes assessments that are carried out in all U.S. schools today. These types of assessments are summative in nature; that is, they provide a summary judgment about, for example, students' learning over some period of time. The goal of summative assessment is to inform external audiences primarily for evaluation, certification, and accountability purposes. Since the federal No Child Left Behind legislation was passed in 2001, summative assessment has certainly received a great deal of publicity in the popular media and has, to a certain degree, swamped the important formative function of assessment.

By focusing on formative assessment, we hope to put assessment back into its rightful place as an integral part of the teaching-learning process. Formative assessment takes place on a continuous basis, is conducted by the teacher, and is intended to inform the teacher and students, rather than an external audience (Shavelson 2006). We view classroom formative assessment as a continuum ranging from informal formative assessment to

formal formative assessment. The position of a particular formative assessment technique on the continuum depends on the amount of planning involved, the formality of technique used, and the nature of the feedback given to students by the teacher. We focus on three important formative assessment techniques—(1) "on-the-fly," (2) planned-for-interaction, and (3) embedded in the curriculum (Figure 2.1) and describe each in turn.

Figure 2.1 Variation in Formative Assessment Practices



On-the-Fly Formative Assessment. On-the-fly formative assessment occurs when "teachable moments" unexpectedly arise in the classroom. For example, teachers circulate between groups to listen in on conversations and make suggestions that give students new ideas to think about. A teacher might overhear a student in a small group investigating sinking and floating say that, as a consequence of an experiment just completed, "Density is a property of the plastic block. It doesn't matter what the mass or volume is, the density stays the same for that kind of plastic." The teacher recognizes that the student has a grasp of what density means for that block, and presents the student with other materials to see if she and her group-mates can generalize the density idea to a new situation. In this way, the teacher challenges the student to test her new idea by having her and her group measure the mass/volume relationships of a new material. Moreover, when satisfied that the students are onto something, the teacher calls for other students to hear what this group found out.

This vision of taking advantage of the "teachable moment" sounds a lot like good teaching, not necessarily *assessment*. This is exactly our point: Teaching and assessment are and should be considered as one and the same.

Rather than teachers planning assessment as a separate event during the class period, on-the-fly assessment is seamless with instruction and is based on the teacher capitalizing on opportunities as they arise to help students to move forward in reaching learning goals.

However, as we learned from our research, such on-the-fly formative assessment and action ("feedback") may be natural for some teachers but difficult for others. Identification of these moments is initially intuitive and then later based on cumulative wisdom of practice. Moreover, even if teachers can identify the moment, they may not have the confidence, techniques, or content knowledge to sufficiently challenge and respond to students.

Planned-for-Interaction Formative Assessment. In contrast, planned-for-interaction formative assessment is deliberate. Teachers plan for and craft ways to get information about the gap between what students know and need to know, rather than use questions just to "keep the show going" during an investigation or whole-class discussion. Consider, for example, teacher questioning—a ubiquitous classroom event. While developing a lesson plan, a teacher can prepare a set of "central questions" that get at the heart of the learning goals for that day's lesson and that have the potential to elicit a wide range of student ideas. For example, these questions may be general ("Why do things sink and float?") or more specific ("What is the relationship between mass and volume in floating objects?" "Can you give me an example of something really heavy that floats? Why do you think it floats?"). At the right moment during class, the teacher poses these questions to the class, and through a discussion the teacher learns what students know and allows different ideas to be presented and discussed. In this example, the teacher planned the assessment prompt in advance rather than waiting for unexpected opportunities to arise. Although not every student in class may respond to each question, the information gained from the students' responses allows the teacher to act on the information collected by fine-tuning instruction or intervening with individual students.

Embedded-in-the-Curriculum Formative Assessment. Alternatively, teachers or curriculum developers may embed more formal assessments ahead of time in the ongoing curriculum to intentionally create "teachable moments." These assessments are embedded after junctures or joints in a

unit where an important goal should have been reached before going on to the next lesson. Embedded assessments inform the teacher about what students currently know and what they still need to learn (i.e., "the gap") so that teachers can provide timely feedback to students.

In their simplest forms, formal formative assessments are designed to provide information on important goals that students should have reached at critical joints in a unit before going onto the next lesson. In their advanced forms, formative assessments are based on a developmental progression of the ideas students have about a particular topic (such as why things sink and float). In contrast to the other two types of formative assessment, embedded assessments are more sophisticated because they are designed to collect critical information about student learning at the same time. The main difference between planned-for and embedded formative assessment is in the designer. Whereas planned-for assessment is usually done by the teacher as a part of the lesson-planning process, embedded assessments are usually designed by curriculum and assessment developers working with experienced teachers.

Embedded formative assessments are valuable teaching tools for at least four reasons. First, they are consistent with curriculum developers' understanding of the curriculum and are therefore consistent with instructional goals. Second, assessment developers contribute technical expertise that increases the quality of the assessments. Third, the involvement of experienced teachers in developing embedded assessments means that they are practical and based on the wisdom of practice. And fourth, embedded assessments provide thoughtful, curriculum-aligned, and valid ways of determining what students know, rather than leave the burden of planning on the teacher.

Formal embedded assessments come "ready-to-use" as part of a preexisting curriculum, and instructional decisions made from them may improve students' learning. Therefore, in our study, we sought to learn whether embedded formative assessments actually helped teachers close the learning gaps in their classrooms.

Potential and Challenges

Formative assessment is a potentially powerful teaching idea embodying knowledge and skills for creating and capitalizing on teachable moments. In the context of science education, formative assessment links teaching

and learning in the service of building students' understanding of the natural world and of how the methods of science justify knowledge claims. In using formative assessments, we sought to move students from naive conceptions of the natural world to scientifically justifiable conceptions ("conceptual change"). To change their conceptions, students need to link what they find out through inquiry investigations to their current conceptions of the natural world and to change those conceptions when their evidence does not fit their "theory." Formative assessment's critical characteristic, then, lies in identifying learning gaps and providing immediate feedback to students that helps them close gaps.

This said, many teachers are in some ways skeptical about incorporating formative assessment substantively into their teaching practice, even when they know that it is important. Teachers have many questions about their role in formative assessment, and for good reason. For example, formative assessment creates a conflict with the teacher's traditional grade-giving role in summative assessment. How can the teacher on the one hand ask students to lay bare their understanding of a concept and at the same time have the responsibility for giving the student a grade? In other cases, teachers may have only experienced *summative* assessment when they were students themselves, or in their teacher education programs. Consequently, they may not have personal experience with the ways that *formative* assessment can improve the quality of teaching and learning. Other questions arise as well. Should teachers really change their beliefs about their role as assessors? Why should teachers change their practices to accommodate a yet unproven teaching technique? Will our emphasis on formative assessment eventually fade away as have other reform techniques?

Clearly, teachers' skepticism is appropriate; part of the science education researcher's role is to test out new (or not so new) techniques to see if they stand up to scientific scrutiny. To this end, our team designed and conducted a study that put formative embedded assessment to the test.

Embedding Formative Assessment in a Science Curriculum

Our study of formative embedded assessment addressed two central research purposes: first, to learn how to build and embed formative assessments in science curricula and, second, to examine the impact of formative assessments on students' learning, motivation, and conceptual change.

Building and Embedding Formative Assessments in Science Curricula

As noted above, we sought to move students from naive conceptions of the natural world to scientifically justifiable ones. To this end, we wanted students to link what they were finding out through investigations to their conceptions about the natural world. The intent was for students to change those conceptions when their evidence didn't fit their "theory."

We embedded formative assessments in the Foundational Approaches in Science Teaching (FAST) curriculum unit on the properties of matter—more specifically, buoyancy (Pottenger and Young 1992). As a first step, we identified the *goals* for the unit. The main goal was for students to develop, through a series of inquiry investigations, a relative density-based explanation for sinking and floating (or, as we came to call it during the study, "Why things sink and float" or "WTSF"). We then worked from the goals backward to the beginning of the unit, identifying key junctures between lessons ("investigations") where important goals needed to be met. We then inserted assessments to provide information about student performance.

Despite our well-conceived plans, in the end, a seemingly straightforward process of developing formative assessments was anything but straightforward. We made some wrong turns and learned from our mistakes.

Pilot Study: From Embedded Formative Assessments to Reflective Lessons

Our basic idea was to develop and embed formative assessments where the "rubber hit the road"—that is, at critical curricular joints where students' conceptual understanding was expected to develop from a simple level to a more sophisticated one. In this way, teachers would know whether students were advancing in their knowledge as the curriculum progressed. We expected that assessments embedded at the critical joints would provide timely information to (a) help teachers and students locate the levels of students' understanding, (b) determine whether students had reached the desired level, (c) diagnose what students still needed to improve, and (d) help students move to the next level.

At each critical joint, we created a set of assessments designed to tap different kinds of knowledge that students should construct in learning about sinking and floating. There were facts (e.g., density is mass per unit volume—*declarative knowledge*) and procedures (e.g., using a balance scale to measure the mass of an object—*procedural knowledge*). But most impor-

tant, and often implicit in curricula, was the use of this declarative and procedural knowledge in inquiry science to build a model or mini-theory of *why* things sink and float (e.g., a model of relative densities—*schematic knowledge*). Consequently, we embedded assessments of these types of knowledge at four natural joints in a 10-week unit on buoyancy. The assessments served to focus teaching on different aspects of learning about mass, volume, density, and relative density. Feedback on performance focused on problematic areas revealed by the assessments.

In order to embed assessments that were based on research and that could identify in a valid and reliable way what students know, we created four extensive assessment "suites" (combinations of individual assessments—graphing, short answer, POE [predict-observe-explain], and PO [predict and observe]). These assessments covered the declarative, procedural, and schematic knowledge underlying buoyancy. Each suite included multiple-choice (with space for students to justify their selections) and short-answer questions that tapped all three types of knowledge. We also included a substantial combination of concept maps (structure of declarative knowledge), performance assessments (procedural and schematic knowledge), predict-observe-explain assessments based on lab demonstrations (schematic knowledge), and/or "passports" verifying hands-on procedural skills (e.g., measuring an object's mass).

Three brave teachers volunteered to try out this extensive battery of embedded assessments in a pilot study. After the completion of the pilot study, the teachers warned us that the original formative assessments were too time-consuming and the amount of information obtained from them was overwhelming. Our lead pilot-study teacher, who was also a member of our assessment team, gently pointed out the problems that pilot-study teachers faced using our assessment suites. She suggested that perhaps there could be only a few assessments that directly led to a single, coherent goal, such as knowing *why things sink and float*. She pointed out that FAST provided ample opportunity for teachers to observe and provide feedback to students on their declarative and procedural knowledge. She urged us to focus on schematic knowledge and on students' developing an accurate mental model of why things sink and float in the assessment suite.

Moreover, Lucks (2003) viewed and analyzed videotapes of the pilot study teachers using the assessment suites. She found that our teachers were treating the "embedded assessments" more as external tests that were some-

thing apart from the curriculum—in other words, as *summative* assessment—rather than using the formative assessments as a way to find out what the students were learning. Thus, the teachers treated the new assessments like any other test that they were required to give to the students during the year, rather than as opportunities to increase their students' learning.

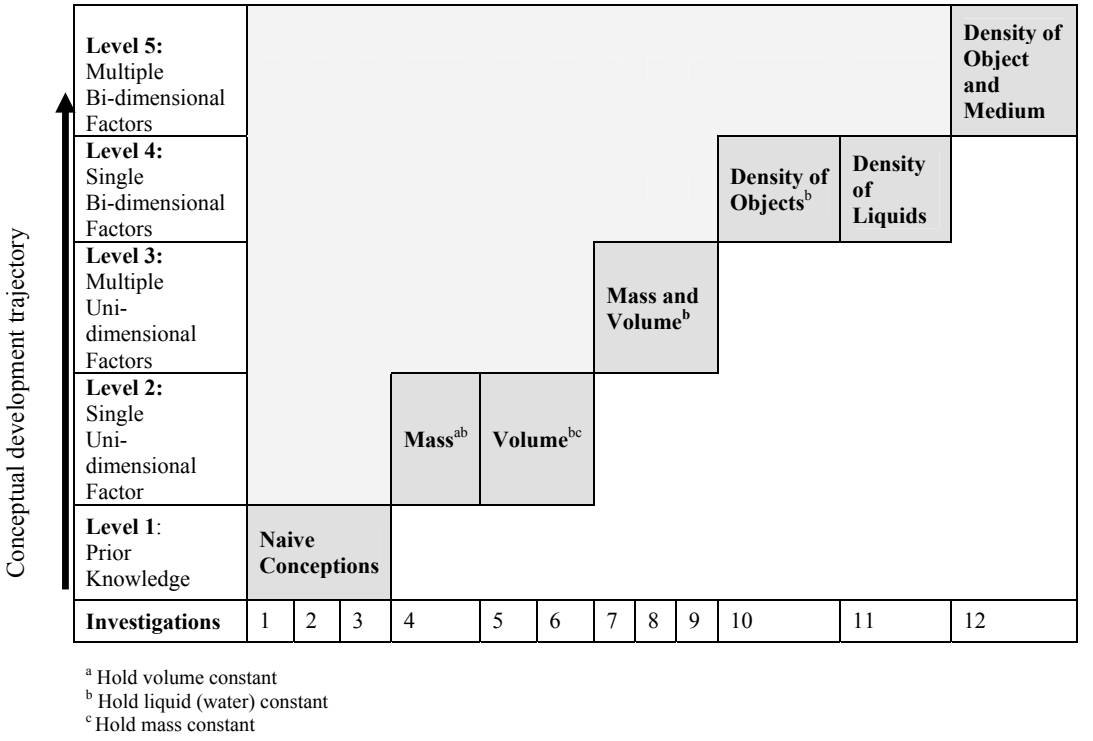
Based on the thoughtful feedback we received from the teachers and the researcher, we revised our initial embedded assessments, greatly reducing their numbers and focusing in on the overarching goal of explaining "why things sink and float." Afterward, when talking with teachers, we no longer spoke of embedded assessments, which we thought would trigger their stereotypes about assessments. Instead, we started calling them "Reflective Lessons" to emphasize their function as a component of the teaching and learning process.

The New Generation of Formative Embedded Assessments: The Reflective Lessons

A second look at the FAST unit and the information collected during the pilot study led us to a developmental progression of student ideas, which then became the basis for redesigning the original embedded assessment suites into Reflective Lessons (Figure 2.2, p. 30). This progression was aligned to the unit and based on different conceptions students have as they develop an understanding of sinking and floating. These conceptions develop from naive (e.g., "things with holes in them will sink") to scientifically justifiable conceptions (e.g., "sinking and floating depend on the relative densities of the object and the medium supporting the object").

Although Figure 2.2 may appear quite complicated, the ideas behind it are straightforward and consistent with students' different ideas about sinking and floating. Before instruction, students have all different kinds of ideas about sinking and floating, such as that heavy things sink, flat things float, things with air in them float. We would place these ideas at Level 1 or "Naive Conceptions." As students progress through the unit, they complete investigations that apply either mass or volume to sinking and floating; that is, a single uni-dimensional factor (Level 2), holding all else constant. Next, students simultaneously apply mass and volume, or multiple uni-dimensional factors, to explain sinking and floating (Level 3). Afterward, students integrate mass and volume into density, a single bi-dimensional factor, in their explanations (Level 4). Finally, students consider

Figure 2.2 Conceptual Development for Understanding Why Things Sink and Float



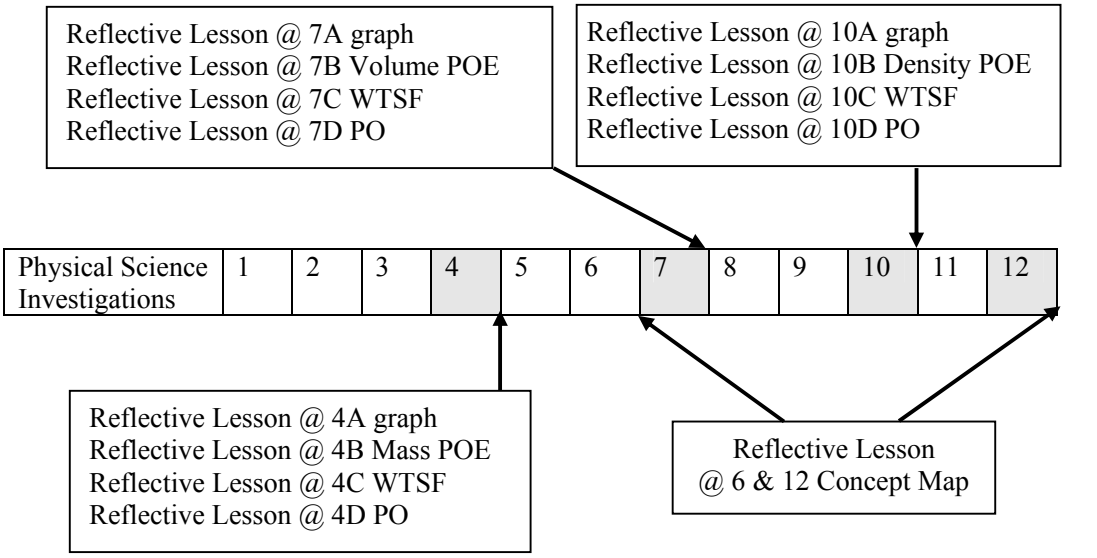
[a] Hold volume constant
[b] Hold liquid (water) constant
[c] Hold mass constant

the object's density and the liquid's density, or multiple bi-dimensional factors (Level 5), in their explanations (Yin 2005).

The final Reflective Lesson suites are shown at their critical junctures in Figure 2.3. Two types of Reflective Lessons were embedded in the unit. Each of the type one Reflective Lessons included a sequence of the following activities: (a) graphing and interpreting evidence and drawing conclusions about WTSF ("Why things sink or float"), (b) applying knowledge learned to predict and explain what would happen in a new situation (Predict, Ob-

serve, Explain), (c) writing a brief explanation about why things sink and float, and (d) predicting and observing a surprise phenomenon to introduce the next set of lessons. The second type of Reflective Lesson was concept mapping, which encouraged students to make connections between the concepts they learned.

Figure 2.3 Reflective Lessons and Junctures at Which They Were Embedded in the Unit



Notes: POE = predict, observe, explain; WTSF = why things sink or float; PO = predict and observe

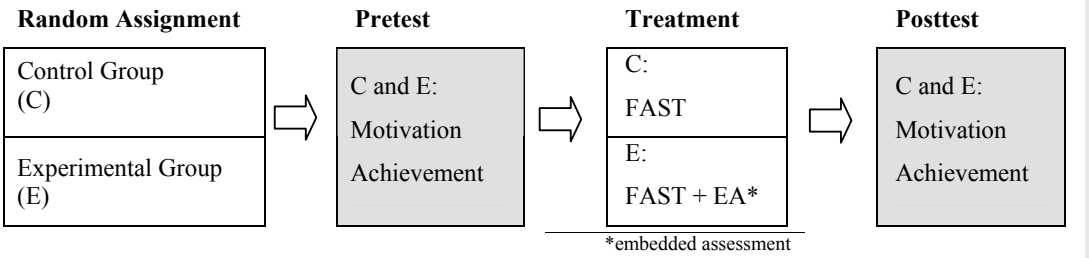
The Reflective Lessons were designed to enable teachers to (a) elicit students' conceptions, (b) encourage communication of ideas, (c) encourage argumentation (comparing, contrasting, and discussing students' conceptions), and (d) reflect with students about their conceptions. In this way, teachers could guide students along a developmental trajectory that they had in hand from naive conceptions of sinking and floating to more scientifically justifiable ones (Figure 2.2).

The Experimental Study

To test whether the final Reflective Lessons could help students improve learning, motivation, and conceptual change, we conducted a small experiment. We randomly assigned 12 teachers to teach either the regular inquiry curriculum (control group—6 teachers) or the curriculum with the Reflective Lessons included (experimental group—6 teachers). Teachers in the experimental group attended a training workshop with the researchers, curriculum developers, and one of the pilot teachers. During the training, teachers participated in the Reflective Lessons as students, talked about the process of the lesson, and then practiced teaching the Reflective Lessons themselves with lab school students. Teachers in the control group also attended a training workshop that oriented them to the study and invited them to share their assessment practices, among other things.

In the study, we gave pretests and posttests to the students in both groups. We examined the effect of the Reflective Lessons by comparing improvement made by the two groups, regarding students' motivation, achievement, and conceptions of sinking and floating (Figure 2.4) (Yin 2005).

Figure 2.4 Schematic of the Research Design



Since the Reflective Lessons integrated formative assessment ideas, curriculum goals, and teachers' input, we expected that students in the experimental group would benefit from the Reflective Lessons and show higher learning gains than the control group. To our surprise, our findings did not support this conjecture. We found no statistically significant differences between average performance in the control and experimental groups. That

is, students in the experimental group and control group did not differ, on average, on motivation, learning, or conceptual change. This finding persisted even after we accounted for differences among students' achievement and motivation before the study began.

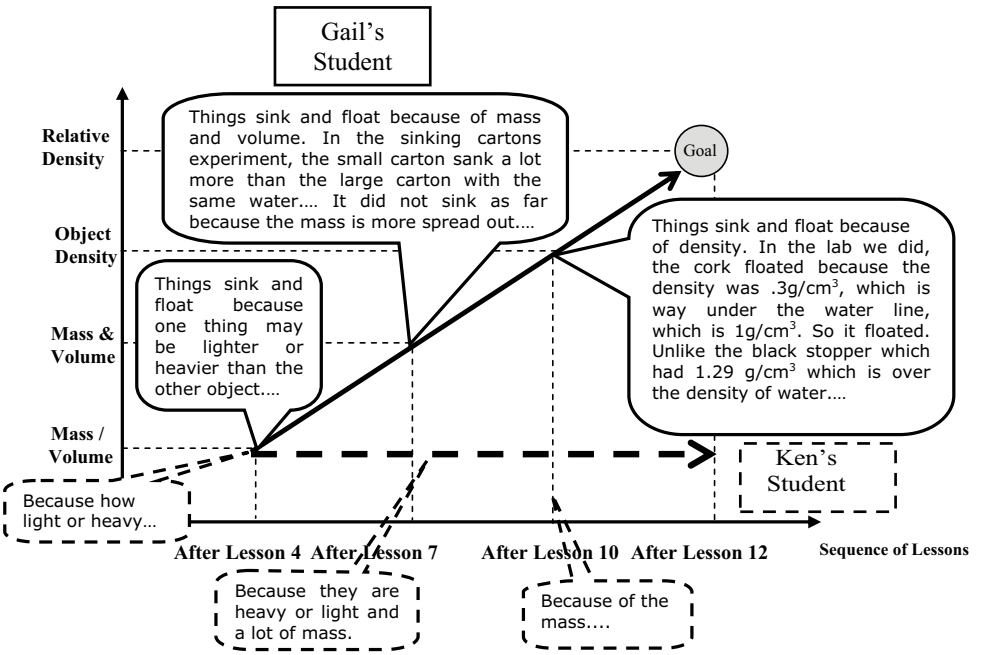
Despite the fact that the study did not come out as expected, we learned a lot about how teachers actually used the Reflective Lessons in their classrooms. In each group, teachers varied substantially in producing differences in students' motivation, learning, and conceptual change. In viewing classroom videos we found that although the Reflective Lessons (embedded assessments) were implemented by teachers in the experimental group, not all the teachers used them effectively to give students feedback or modify teaching and learning (Ruiz-Primo and Furtak 2006, 2007). That is, among the teachers in the experimental group, those teachers whose students had higher learning gains relied more on the other two types of assessment techniques—on-the-fly and planned-for-interaction assessment—rather than on the Reflective Lessons.

To give an idea of the differences among teachers, let us consider two teachers in the experimental group, Gail and Ken.[1] Gail took an active role in using the Reflective Lessons with her students. She would build knowledge with students by challenging their ideas, asking them for empirical evidence to justify their ideas, and making clear how a model of sinking and floating was emerging. The Reflective Lessons created teachable moments for her, which she then took advantage of with informal assessment techniques. Ken, in contrast, relied on the Reflective Lessons themselves to help the students learn and looked at the activities as discovery learning; that is, he depended on the students to develop their own understandings with limited teacher intervention (Furtak 2006). He reasoned that it was not his role to act on the students' ideas about sinking and floating and to guide the students or tell them the answers; rather it was up to students to discover for themselves why things sink and float.

In Figure 2.5, page 34, we see the developmental trajectory for a typical student from Gail's class and another from Ken's. While Gail's student progressed along the trajectory, Ken's student held to her original explanation. The achievement test scores for the two students reflected the differences

[1]These names are pseudonyms. We use male and female names for writing ease (e.g., to avoid he/she, his/her). We did not find gender differences in teaching effects in our study.

Figure 2.5 Development of Understanding of Why Things Sink and Float in Two Experimental Teachers' (Gail's And Ken's) Students



in learning (Gail's student: pretest 15 and posttest 36; Ken's student: 23 and 23, respectively) (Yin 2005).

Concluding Comments

As we know, when any new reform idea comes along, there is a lot of hype. Moreover, teachers are expected to pick up the new "tools" and implement the ideas perfectly on the first try, after they have been trained (briefly!) to do so. Even though we worked intensely with our experimental teachers to learn how to use Reflective Lessons and provided follow-up during the experiment, the kinds of knowledge, belief, and practice changes we wanted to bring about—conceptual changes—needed much more time. Those