On the Fidelity of Implementing Embedded Formative Assessments and Its Relation to Student Learning

Erin Marie Furtak \textsuperscript{a}; Maria Araceli Ruiz-Primo \textsuperscript{b}; Jonathan T. Shemwell \textsuperscript{c}; Carlos C. Ayala \textsuperscript{1}; Paul R. Brandon \textsuperscript{e}; Richard J. Shavelson \textsuperscript{c}; Yue Yin \textsuperscript{g}

\textsuperscript{a} Max Planck Institute for Human Development, \textsuperscript{b} University of Colorado at Denver and Health Sciences Center, \textsuperscript{c} School of Education, Stanford University, \textsuperscript{1} School of Education, Sonoma State University, \textsuperscript{e} College of Education, University of Hawaii, \textsuperscript{g} College of Education, University of Illinois, Chicago \textsuperscript{g} College of Education, University of Hawaii, Manoa

Online Publication Date: 01 October 2008

To cite this Article Furtak, Erin Marie, Ruiz-Primo, Maria Araceli, Shemwell, Jonathan T., Ayala 1, Carlos C., Brandon, Paul R., Shavelson, Richard J. and Yin, Yue(2008)'On the Fidelity of Implementing Embedded Formative Assessments and Its Relation to Student Learning'.Applied Measurement in Education,21:4,360 — 389

To link to this Article DOI: 10.1080/08957340802347852

URL: http://dx.doi.org/10.1080/08957340802347852

Please scroll down for article

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

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

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

Erin Marie Furtak
Max Planck Institute for Human Development

Maria Araceli Ruiz-Primo
University of Colorado at Denver and Health Sciences Center

Jonathan T. Shemwell
School of Education
Stanford University

Carlos C. Ayala
School of Education
Sonoma State University

Paul R. Brandon
College of Education
University of Hawaii

Richard J. Shavelson
School of Education
Stanford University

Yue Yin
College of Education
University of Illinois at Chicago

College of Education
University of Hawaii at Manoa

1Order of authors at this point is alphabetical.
Correspondence should be addressed to Erin Marie Furtak, School of Education, University of Colorado at Boulder, Boulder, CO 80309. E-mail: erin.furtak@colorado.edu
Given the current emphasis on conducting high-quality experimental studies, it is becoming increasingly important for researchers to accompany their studies with evaluations of the fidelity of implementation of the experimental treatments. This article compares the form and extent of an experimental treatment to student learning. The study involved six middle-school physical science teachers and their students. Videotaped lessons were coded according to their alignment with the treatment’s intended structure and processes; results identified a 0.71 correlation between teachers’ enactment of formative assessment and student learning. The article suggests that the results of the study may have been due, at least in part, to the failure of many of the experimental-group teachers to implement the treatment as intended.

INTRODUCTION

Teachers in American public schools have a wide range of academic specializations and abilities, take a variety of paths to certification, and bring their own backgrounds, beliefs, and experiences into their classrooms (Richardson, 1996; Zumwalt & Craig, 2005). Furthermore, the context of every classroom varies with the student population, the conditions of the school, the community, and the district and state in which a teacher works. This diversity means that providing any curriculum to six different teachers across multiple schools and states will result, to a certain degree, in six variations on what curriculum developers intended. A similar contention can be made with respect to instructional treatments. In the case of our project, the effectiveness of embedded formative assessments depended not only on the quality of the assessment prompts as they were designed (Furtak & Ruiz-Primo, 2007), but also on how they were implemented. This means that in order to draw valid conclusions relating to the potential of formative assessments to improve students’ learning, it is critical to know whether teachers implemented the formative assessments as intended by their designers.

We therefore explored in this study what teachers actually did with embedded formative assessments, called “Reflective Lessons.” In the absence of this information, it would be difficult to determine whether the results observed in the project (see Yin et al., this issue) can be attributed to an absence of a formative-assessment treatment effect, a poor conceptualization of formative assessments in this study, or to an implementation that not only varied between teachers, but also strayed considerably from what had been intended by the assessment designers. By looking at implementation, we move beyond the mere design of the instructional treatment to compare the form and extent of the treatment teachers actually delivered to the observed learning gains of their students. More importantly, we identify shortcomings in the design and implementation of embedded assessments that must be overcome to be effective instruments for learning in the classroom.
In this article, we provide one possible model for examining fidelity of treatment implementation. Using this model as an analytic lens, we focus on the following research questions:

1. Did the teachers implement the critical characteristics of the embedded assessments as envisioned by the Assessment Development Team (ADT) and as described in the Teacher’s Guide to the Reflective Lessons?
2. Was implementation fidelity related to students’ learning?

In what follows we provide the framework we used to approach the study of the implementation of the embedded formative assessments. We then describe the methodological characteristics of the study. Finally, we present the results and the lessons learned from them.

AN APPROACH FOR MEASURING FIDELITY OF IMPLEMENTATION

Fidelity of implementation is generally considered to be a way of determining the alignment between the implementation of a treatment and its original design. However, there is no clear consensus on what exactly constitutes fidelity of implementation, and empirical evidence on the relationship between fidelity of implementation and program success is limited (see Dane & Schneider, 1998; Dusenbury, Brannigan, Falco, & Hansen, 2003; Ruiz-Primo, 2005). In this article, we intend to contribute evidence that links fidelity of implementation to treatment effectiveness in the service of better understanding the results of the project.

The process of designing, implementing, and measuring a treatment can be divided into three categories for analysis: the intended, enacted, and achieved effectiveness of the treatment (McKnight et al., 1987; Ruiz-Primo, 2005). In this article, we attempt to connect these categories. To do so, we focus on two aspects of fidelity of implementation as defined by Dane and Schneider (1998) that deal with how well the enacted curriculum matches the intended. The first aspect is adherence, or the extent to which specified components of a program, curriculum, or treatment are delivered as prescribed; and the second is quality of delivery, or the extent to which teachers approach a theoretical ideal in terms of prescribed processes of a program, curriculum, or treatment (Figure 1).

Measuring fidelity of implementation must begin by identifying the critical treatment characteristics that are supposed to achieve its effects (Bauman,

---

1We acknowledge that this approach has been used in studying curricula rather than treatments, but we believe the strategy equally applies to the study of treatments.
Stein, & Ireys, 1991; Ruiz-Primo, 2005). Clear specification of what the treatment entails is necessary to ensure that the active ingredients critical to its success are being delivered (Moncher & Prinz, 1991). We identified the critical characteristics based on adherence and quality of delivery. While adherence refers to the implementation of the structure of the treatment, quality describes the implementation’s fidelity to the process of the treatment (Mowbray, Holter, Teague, & Bybee, 2003; Ruiz-Primo, 2005). Therefore, to evaluate the extent to which the teachers in the experimental group adhered to the Reflective Lessons as intended, we focused on the structure and the delivery process of the Reflective Lessons as defined in the intended treatment. Finally, we used the results from the comparison of the intended and enacted treatments to guide our interpretation of the achieved curriculum.

DEFINING CRITICAL ASPECTS OF THE EMBEDDED ASSESSMENTS

As described in other articles in this issue (e.g., Ayala et al., this issue), the embedded formative assessments are formal prompts inserted into the curriculum that are designed to help teachers check student understanding at key points during instruction and reflect on the next steps needed to move students forward in their learning. However, it is important to go beyond defining the goals of the embedded assessments (e.g., reduce the gap between where students are and where they should be) or describing general requirements for their administration (e.g., formative assessments require three days to implement) so that we may define the critical operational features or aspects of the embedded assessments. Defining the critical aspects requires a careful analysis of the treatment at hand; in this case, a careful analysis of the envisioned or intended structure and process of the embedded assessments and their implementation. At times, the critical aspects of the embedded assessments may seem indistinguishable from teaching behaviors; this reflects the intent of the assessments as teaching tools that would help teachers elicit students’ ideas in normal lessons that are not apart from everyday teaching.
Adherence: Treatment Structure

In order to measure teachers’ adherence to the Reflective Lessons as the ADT intended, we first determined the critical aspects of the treatment using the FAST Teacher’s Guide to the Reflective Lessons (2003, referred to as Guide). This Guide was the primary source of information about the treatment and was used to design and carry out the summer training program with the experimental teachers.

Two types of embedded formative assessments were used in the study (e.g., Ayala et al., this issue). They varied based on the type of assessment prompt used and the structure of implementation. Reflective Lessons Type I consisted of four formative assessment prompts (Graph, Predict-Observe-Explain, Short-Answer, and Predict-Observe) to assess students’ conceptions and/or mental models around why things sink and float, and to support students in fashioning increasingly coherent and evidence-based explanations of the phenomena. Reflective Lessons Type II were concept maps and focused on checking students’ progress in understanding key concepts in the unit. Each type of Reflective Lesson had a different structure of implementation, and thus presents different critical program components to be measured.

In the Reflective Lesson Type I, formative assessment prompts were intended to build on each other; therefore, it was expected that all prompts would be implemented in the sequence prescribed. Furthermore, teachers were to intersperse discussions within the sequence of written prompts. In the case of Predict-Observe-Explain, teachers were provided with three possible sequences they might use to mix written work, class discussion, and teacher demonstrations. Based on this information about the structure of the Reflective Lessons Type I, we considered three aspects of the treatment as critical to their effectiveness: (1) implementation of each assessment prompt, (2) sequence in which they were to be implemented, and (3) placement of discussions between written prompts. We also identified a fourth component, the amount of time teachers would take to implement the prompts, not as being critical to the effectiveness of the Reflective Lessons, but as an important piece of information about the feasibility of using the embedded assessments. The ADT envisioned the Reflective Lessons Type I to be carried out across two to three 45-minute class periods, although the exact amount of time teachers used was not discussed at the minute level in the introductory workshop. Figure 2 illustrates the implementation structure and critical components of the Reflective Lessons Type I.

The Guide also suggested an order for the Predict-Observe-Explain assessment, related to sequencing, which we expanded to include not only the sequence between prompts, but also sequence within prompts. Since teachers were given three options for the sequence of activities in carrying out the Predict-Observe-Explain assessment, we viewed this aspect as non-critical. Each sequence
involves the students recording their predictions and reasons, the teacher collecting, clustering, and posting those predictions and reasons, and asking students to write their observations and explanations. The sequences differed in the placement of discussions, and at which point the students are asked to write their observations and explanations.

In the Reflective Lesson Type II, the Guide specified that students should create concept maps as individuals, and then combine their best ideas into a small-group concept map. The Reflective Lesson Type II was intended to be carried out in one class period. Therefore, three of the four aspects from Reflective Lessons Type I were also identified for this type of Reflective Lesson: (1) implementation of prompts (i.e., individual and group concept maps), (2) implementation in sequence, and (4) timing. Teachers were given one possible sequence for implementing the Concept Maps: to begin by training students in the procedure for making the map, then having students make a map working individually, a map in a small group, and then constructing a concept map as a class.

Quality of Delivery: Treatment Processes

The evaluation of the quality of the treatment processes of the Reflective Lessons focused on the teaching strategies (as they are called in the Guide) conceived by the ADT and the summer teacher training. These strategies were developed to be consistent with models of formative assessment in scientific inquiry settings. In
this article, we define the critical treatment processes of our Reflective Lessons in terms of the two major formative assessment processes they embodied: making students’ thinking visible and advancing students in their learning. We divided the first process, making students’ thinking visible, into two critical aspects: (1) eliciting (publicly) student conceptions about sinking and floating and (2) tracking and clustering these conceptions in relation to each other and to our target learning trajectory. Advancing students’ learning of the program content included three more critical aspects: (3) helping students provide reasons for their explanations; (4) encouraging argumentation; and (5) helping students base their claims on evidence collected from in-class investigations. These processes are described in more detail later in the article.

Since formative assessment assumes that teachers’ instructional actions must be based on what students currently know (e.g., National Research Council, 2001a), a fundamental element of its enactment is eliciting and making public students’ conceptions. In our project, teachers were provided with lists of strategies to help elicit students’ ideas, including asking students to come to a consensus at their table, facilitating student presentations, taking votes, and simple questioning in whole-class, small group, and individual settings.

Because students can quickly produce a wide range of conflicting or redundant ideas in scientific inquiry settings, teachers can monitor students’ ideas in formative assessment settings by recording or making them visible in some way. The Guide specifically asked teachers to track students’ conceptions and present them in a visual manner, such as writing students’ ideas on the board, tallying votes for predictions, or recording ideas on pieces of paper that could be moved around and compared. In addition, teachers were specifically asked to cluster students’ conceptions, consolidating similar ideas and summarizing them into central ideas. For instance, a teacher might collapse ideas like “flat things float more easily” and “boats might be heavy, but they still float” into a more general statement such as “shape matters.”

Another important element to formative assessment teaching strategies is for students to communicate their ideas to each other, and to provide reasons, evidence, and explanations for their ideas (Black & Wiliam, 1998). The teacher’s role therefore is to promote reasoning, by asking students to provide explanations and justifications, probing for deeper meaning, and comparing/contrasting student ideas (Ruiz-Primo & Furtak, 2006; 2007). A focus of the training program and the Guide was to train teachers to push students to clarify and support their ideas.

In the context of formative assessment, argumentation can serve the function of self- and peer-assessment, where students listen to the ideas of others, consider supporting evidence, and progress to higher levels of understanding (Sadler, 1989). Arguing scientific ideas is also fundamental to the practice of scientific inquiry, both in the classroom and in the field of science (e.g., AAAS, 1990;
Therefore, in the Guide and training, teachers were encouraged (and were given opportunities to practice during the summer training) to promote student-to-student discussions and debate rather than merely responding to questions posed by the teacher. This argumentation was intended to provide students immediate feedback about their conceptions as they reflected on how evidence could be used to support their claims.

Finally, the training and Guide emphasized that teachers should encourage students to provide evidence for their ideas, so that this evidence might be evaluated and used to revise knowledge claims. Evidence-based reasoning is a cornerstone of effective formative assessment practice in the context of scientific inquiry (National Research Council, 1996; 2001b; Duschl, 2003). To capture the scientific inquiry nature of this instructional transaction, we also created a component named student use of evidence-based reasoning to capture whether or not students were citing evidence from the investigation they completed in class, and whether or not this evidence was then used to refine, develop, and support universal explanations for sinking and floating. Table 1 provides a summary of the analytic framework to determine the enacted treatment for the implementation study.

### METHOD

This section provides information about the six Experimental Group teachers and their classes, data collection and analysis procedures.

#### Participants

The six teachers who were randomly assigned to the Experimental Group represent various backgrounds and levels of experience; not all of the teachers in the

<table>
<thead>
<tr>
<th>Intended Treatment</th>
<th>Critical and Non-Critical Aspects</th>
</tr>
</thead>
</table>
| Adherence: Treatment Structures | (1) Implementation of all prompts  
(2) Sequence within and between prompts  
(3) Placement of discussions  
(4) Timing within and between prompts |
| Quality of Delivery: Treatment Processes | (1) Eliciting student conceptions  
(2) Tracking and clustering student conceptions  
(3) Asking students to provide reasons for their explanations  
(4) Students argue ideas and evidence  
(5) Students provide evidence for their claims |

1993; Newton & Osborne, 1999; Osborne, Erduran, & Simon, 2004).
study had post-secondary background in science, and several were close to the beginning of their teaching careers (see Shavelson et al., this issue). The curriculum was implemented in sixth grade in some districts, and seventh grade in others. Class sizes ranged from 20 to 31, and class-period length also varied.

Sources of Information

**Intended treatment**

The *Guide* served as the major source of information about the intended treatment, as it reflected both the intentions of the ADT, as well as the summer training carried out with the experimental group teachers (see Ayala et al., this issue).

**Enacted treatment**

Evidence about how the Reflective Lessons were implemented came from classroom videotapes of each teacher’s focus class (for information on how focus classes were selected, see Yin et al., this issue).

**Achieved treatment**

Students’ responses to the 38 items that appeared on the pre- and posttests of achievement were used as evidence to link the quality of the treatment implementation on the effectiveness of the treatment (see Yin et al., this issue).

Data Collection

**Videotapes**

Because teachers were distributed across the country, it was not feasible to have staff videotape the implementation of the Reflective Lessons. Rather, teachers in both the control and experimental groups in the original study were shown how to set up and videotape lessons; these videotapes served as the primary data source for this study. Because the intent of this article is to determine the fidelity of implementation of the formative-assessment treatment, and not to compare the instructional practices of teachers in the control versus experimental groups, only the videotapes from the experimental teachers were separated for the present analysis.

As part of the summer training workshop, all teachers in the experimental group were shown a videotape explaining how to place the camera in the classroom and how to familiarize students with the equipment. Each teacher was then given a Canon ZR60 Digital Video Camcorder, a tripod, a lapel microphone with
pocket transmitter, and an ample supply of mini-DV tapes and batteries for operating the cameras. Teachers then practiced setting up and operating the camera under the guidance of members of the Stanford Education Assessment Laboratory (SEAL) members. Each teacher then placed the camera at the back of their classroom so that the teacher and some students could be observed. Teachers videotaped each day that they taught investigations from the first 12 investigations of FAST in which the assessments were embedded, as well as all of the Reflective Lessons, which were embedded after lessons 4, 6/7, 10/11 (see Ayala et al., this issue, for more information about the design and placement of the Reflective Lesson prompts).

Each week, teachers submitted their videotapes in pre-addressed, stamped envelopes. SEAL staff then logged the videotapes according to date and lesson taught into a database, numbering each tape sequentially so that they could be easily kept in order. An outside contractor transcribed the videotapes and converted them to RealPlayer files so that the videos could be copied and viewed.

Each videotape marked by the teachers as containing a Reflective Lesson was separated for analysis. Because teachers took different amounts of time to teach the sinking and floating unit in which the assessments were embedded, the total number of Reflective Lesson videotapes collected from each teacher varied greatly from 8–16 videotapes of 424–825 minutes duration.

Although most of the teachers submitted videotapes of most of the prompts, Becca did not submit videotapes of four prompts—more than one-fourth of the Reflective Lessons. The other exception was Andy, who did not submit any videotapes of Concept Maps. Of the tapes received, about 5% had poor or no sound and were not further analyzed.

**Pre- and posttests of student achievement**

The pre–posttests of achievement included proximal and distal multiple-choice and short-answer items addressing the concepts of mass, volume, density, relative density and sinking and floating, and graph interpretation. Thus student performance on the pre- and posttest can be identified as a measure of various aspects of students’ conceptual understanding. More information on the validity and reliability of the 38 common pre–posttest multiple-choice achievement items can be found in Shavelson et al. (this issue) and Yin et al. (this issue).

Pretests were administered to students at the beginning of the students’ science course. In most cases, this testing occurred in the fall (because Carol’s science class did not start until January, pretests were administered to her students after the winter break). Posttests were then administered as closely as possible to the last day of instruction.
Data Analysis

Information about the coding system used to determine the fidelity of implementation is presented first in what follows. Then, the means by which information from this coding system was used to determine the Intended treatment from coding the teacher guide, the Enacted treatment by coding the videotapes, and the Achieved treatment in terms of analysis of student learning is explained.

Fidelity of implementation coding system

We designed a coding protocol to capture in the most direct way possible each teacher’s alignment with the intended treatment—the critical aspects falling under the headings treatment structures and quality of delivery (Table 1). These critical aspects were operationalized into one or more categories intended to provide measures that capture the extent to which each teacher enacted the treatment as intended (Table 2). The first two codes focus on the class organization suggested in the Guide to promote certain processes. For example, the more time spent on teacher talk/task setting, the less time was available for other important processes (e.g., argumentation). The next three codes focus on events that were expected to be observed in an appropriate treatment process.

Intended treatment: Coding the teacher guide

The aforementioned codes were used, when relevant, to identify what might be called the “ideal implementation profile,” that is, the ADT’s vision of treatment implementation. Codes were applied to relevant pieces of the Guide; for example, applying the “Focus of Instruction” codes to the treatment implementation sequences described earlier. In other cases, information from the guide was translated directly into the ideal profile of implementation. For example, the Guide’s suggestion that Reflective Lessons should be implemented in 2–3 days was used to interpret the length of time that teachers actually took to implement the lessons.

Enacted treatment: Coding the videotapes

The “Fidelity of Implementation” coding system was then applied to the videotapes of the Reflective Lessons collected by the teachers. One-minute time intervals served as the unit of analysis. Because we wanted to capture all relevant strategies used by each teacher, and to be able to make statements about the duration of different strategies and elements to the Reflective Lessons, coding the videotapes minute-by-minute provided a broad overview of how each teacher used his or her time to implement the lessons.

To begin, all videotapes submitted by each teacher were reviewed, and those that were labeled as containing Reflective Lesson material were verified and designated for analysis. Then a transcript of each video was segmented by minute. In
### TABLE 2
Description of the Categories and Codes Used in Coding Protocol

<table>
<thead>
<tr>
<th>Category</th>
<th>Codes</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatment Structures</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Focus of Instruction</strong></td>
<td>Teacher talk/task setting</td>
<td>Teacher speaks to class without engaging students in discussions, sets Reflective Lesson task, or carries out demonstration</td>
</tr>
<tr>
<td>Individual</td>
<td></td>
<td>Students work individually</td>
</tr>
<tr>
<td>Small Group Talk</td>
<td></td>
<td>Students work together in pairs or small groups</td>
</tr>
<tr>
<td>Whole Class Discussions</td>
<td></td>
<td>Teacher and students engage in a discussion, or teacher works with students’ ideas</td>
</tr>
<tr>
<td><strong>Student Task</strong></td>
<td>Table consensus/survey</td>
<td>Students at each table come to consensus regarding issue, confusion, or conception; or students collect from their groups all conceptions about the question of interest</td>
</tr>
<tr>
<td>Peer review</td>
<td></td>
<td>Students review work of their peers using teacher-provided answer sheet</td>
</tr>
<tr>
<td>Self-review</td>
<td></td>
<td>Students review own work using teacher-provided answer sheet</td>
</tr>
<tr>
<td><strong>Treatment Processes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Making students’ thinking explicit</td>
<td>Collecting</td>
<td>Teacher asks students to share their observations, predictions, hypotheses, evidence, examples, definitions, procedures or to answer simple, yes–no, or fill-in-the-blank questions.</td>
</tr>
<tr>
<td>Class Presentation</td>
<td></td>
<td>Teacher asks students to report in front of class, either individually or in groups</td>
</tr>
<tr>
<td>Voting</td>
<td></td>
<td>Teacher asks students to raise hands to indicate their prediction/explanation/conception; tally captured in some way on board</td>
</tr>
<tr>
<td>Tracking student conceptions</td>
<td>Displaying Students’ ideas (no clustering)</td>
<td>Students’ ideas displayed on board/overhead/papers/posters/etc. without explicitly organizing or grouping. If ideas are displayed and the discussions about that idea continues into next minute, continue coding the idea as displayed until a new idea is discussed</td>
</tr>
<tr>
<td>Clustering Students’ ideas</td>
<td></td>
<td>Teacher actively clusters, categorizes, funnels, or groups students’ ideas, concepts, procedures, or terms (Teacher arranges ideas into groups or categories; can only happen after teacher has collected at least one conception)</td>
</tr>
<tr>
<td>Promoting Reasoning</td>
<td>Asking for students’ explanations, reasoning, or justification</td>
<td>Teacher asks “why” questions which are initial or follow-up queries to elicit students’ thinking and make students’ reasoning explicit. This also includes teacher asking students to provide evidence for claims they have made</td>
</tr>
</tbody>
</table>

(Continued)
some cases, the process of segmenting the videotapes revealed that some of the lessons were not properly marked; that is, these lessons contained other activities not related to the Reflective Lesson implementation. These videotapes were then removed from the sample.

The coding strategy was then applied to each time interval once, with only a single code permitted for each category. The “Focus of Instruction” code was applied first and determined which subsequent codes would be applied or designated as “not applicable”; for instance, the “Student Task” code was only applied when students were working individually or in small groups, and other codes were only applied during whole-class discussions. Table 3 illustrates how codes were applied to four minutes of videotape; coding categories are shown in gray if they were not applied to a particular minute of videotape.

The first three authors coded the videotapes. As a check on first two authors’ familiarity with the data, the third author was not fully aware of the hypotheses for the study. Once satisfactory levels of agreement were reached in reliability training analyses, each rater independently coded 22% of the videotaped lessons. All statistics exceed satisfactory threshold levels (Percent direct inter-rater agreement: Mean = 97%, Min. = 96%, Max. = 99%; Cohen’s Kappa: Mean = .89, Min.
### TABLE 3
Example of Four Coded Minutes

<table>
<thead>
<tr>
<th>Speaker/Dialogue</th>
<th>Coding Category</th>
<th>Codes Applied</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Minute 1</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher:</td>
<td>Focus of Instruction</td>
<td>Teacher Talk/Task Setting</td>
</tr>
<tr>
<td></td>
<td>Focus of Instruction</td>
<td>Teacher Talk/Task Setting</td>
</tr>
<tr>
<td></td>
<td>Student Task</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Making Students’ Thinking Explicit</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tracking Student Conceptions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Promoting Reasoning Argumentation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Role of Evidence</td>
<td></td>
</tr>
<tr>
<td>Friday, we left off with you predicting which straw would sink the furthest, which would have the greatest depth of sinking, and which straw would have the least depth of sinking. So I want you to have that page in front of you. Make sure you have that page in front of you. What I’m going to do now is take a quick survey. So everyone look very carefully at your paper. Look at the part of the table that says greatest depth of sinking and look to see which straw you circled,</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **Minute 2**     |                 |               |
| Teacher:         | Focus of Instruction | Whole Class Discussions |
|                  | Focus of Instruction | Whole Class Discussions |
|                  | Student Task     |               |
|                  | Making Students’ Thinking Explicit |              |
|                  | Tracking Student Conceptions |            |
|                  | Promoting Reasoning Argumentation |          |
|                  | Role of Evidence  |               |
| Look to see which straw number you circled, and then put your head down. You should know what page it is. Look to see on the table which straw you predicted would sink the greatest, keep it in your head and put your head down. Raise your hand if you thought that straw number one would sink the furthest? Your head should be down, Nate. Raise your hand if you think that straw number two would sink the furthest? Raise your hand if you thought straw number three would sink the furthest. (Teacher writes tally on board after each vote), |

| **Minute 3**     |                 |               |
| Teacher:         | Focus of Instruction | Whole Class Discussions |
|                  | Focus of Instruction | Whole Class Discussions |
|                  | Student Task     |               |
|                  | Making Students’ Thinking Explicit |              |
|                  | Tracking Student Conceptions |            |
|                  | Promoting Reasoning Argumentation |          |
| And raise your hand if you thought straw number four would sink the furthest. Okay, raise your heads. These are the results. Zero people thought straw number one would sink the furthest, one person thought straw number two would, one person thought straw number three would, and 20 people thought that straw number four would. Now, we have some interesting things going on here. For those of you who thought that straw number four would sink the furthest, what were some reasons for that?, |

| Role of Evidence  |               | No evidence cited |

(Continued)
Once agreement was established, the remaining videos were divided between two of the three raters and coded independently.

**Achieved treatment: Analysis of student learning**

We tested differences between the six teachers’ students before and after instruction. We also examined evidence about the relationship between the quality of the implementation and student learning.

**RESULTS**

The goals of this component of this study were to determine whether the teachers implemented the critical aspects of the embedded assessments as prescribed by the Guide and to link the quality of implementation to the effectiveness of the formative
assessments to improve student learning. This section is organized to address these goals directly. First, information regarding the Intended and Enacted Treatment will be presented; then the results of student learning—the Achieved Treatment—is presented. Finally, we attempt to link the Intended, Enacted, and Achieved Treatments together to guide interpretation of the Romance Project findings.

**Intended and Enacted Treatment**

For the purpose of clarity, results for the Intended and Enacted Treatment are presented first for adherence to treatment structure, and second for quality of delivery of treatment processes. Each sub-section is subsequently presented according to the aspects of fidelity of implementation identified earlier in the article.

**Measuring adherence to treatment structure:**

(1) **Implementation of all prompts.** Our expectation was that teachers would implement all of the prompts that made up the Reflective Lessons Types I and II. This expectation was reiterated in the summer teacher training and in the *Guide*. Unfortunately, because teachers were responsible for submitting videotapes themselves, we cannot know if we received videotapes of all the lessons that were taught. Given that limitation, we still attempted to determine if we had videotape evidence from all teachers enacting all Reflective Lesson prompts. Diana was the only teacher who submitted videotapes from every prompt, and Aden, Carol, and Robert were missing only one. Andy and Becca were missing three and four lessons, respectively.

(2) **Sequence within and between prompts.** The *Guide* supplied model sequences for the Reflective Lessons Type I and II overall, as well as sequences for the implementation of the Predict-Observe-Explain assessments in particular. Although we emphasized the importance of implementing all prompts in sequence, teachers were provided with several options for the implementation of the Predict-Observe-Explain assessment, which blend multiple teaching strategies and levels of grouping.

With respect to within-prompt sequences, we coded the three suggested Predict-Observe-Explain sequences according to grouping level and strategies. Half the teachers experimented with multiple suggested implementation sequences. The data also show that two of the teachers—Carol and Diana—did not use the suggested sequences at all; instead, these two teachers blended small group work within the sequences provided by the *Guide*.

Turning to the between-prompts level of analysis, we found that although five of the six teachers did not implement all prompts, all six teachers always
implemented the Type I Reflective Lessons in the sequence described by the *Guide*. In contrast, the teachers were less likely to follow the sequence guidelines for the Reflective Lessons Type II. A closer look at the data indicates that teachers with sequences categorized as “other” often skipped some elements of the concept map Reflective Lessons, for example, making a group map. In other cases, teachers varied focus of instruction (e.g., whole-class discussion, small group) more than suggested by the sequences in the teacher guide (e.g., proceeding in order from individual work to small group work to whole-class discussion). Results of these within and between-prompt sequence analyses are shown in Table 4.

**Table 4**

<table>
<thead>
<tr>
<th></th>
<th>Within-Prompt Predict-Observe-Explain</th>
<th>Between-Prompt RL Type II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Followed Sequence</td>
<td>Other</td>
</tr>
<tr>
<td>Aden</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Andy</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Becca</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Carol</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Diana</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Robert</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

(3) **Placement of discussions.** The *Guide* instructed teachers to initiate and promote scientific discussions during the Reflective Lessons. During the summer training, teachers were given opportunities to hold discussions with students during each type of embedded assessment prompt. Figure 3 illustrates the relative percentages of time that each teacher devoted to whole-class discussion, as well as other foci of instruction. Teachers spent an average of 27% of Reflective Lesson class time in whole-class discussion (Min. = 20%, Max. = 35%).

Looking more closely at when these discussions happened, we focused on the discussions that fit within the Reflective Lesson Type I sequence after the Short-Answer and Predict-Observe-Explain prompts. A discussion was viewed to have “taken place” if there was at least one minute of discussion after the Predict-Observe-Explain task was completed, or if more than 1 minute of discussion took place after a period of independent student work following the Short Answer prompt. This standard was applied to the Short Answer because the teachers were encouraged strongly to have discussions of longer duration at this point in the unit. A comparison of where teachers placed discussions is provided in Table 5.

Among the six teachers, none held a discussion following every Predict-Observe-Explain prompt. Results of the Short Answer prompt indicated that
while Carol and Diana never missed a discussion, Robert, Andy, and Becca skipped discussions once, and Aden twice. The fact that many of these Short Answer whole-class discussions were missed is important, because this was the place during which the treatment processes, to be discussed later, were expected to take place.

(4) Timing within and between prompts. According to the Guide, Reflective Lessons Type I were to be implemented in two to three class sessions, assuming a 45-minute class period. Reflective Lessons Type II were to take one class session to implement. According to the sequence of implementation described in the Guide, conservative estimates can be made of how long each prompt might take to implement. Table 6 combines these estimates with each teacher’s average duration of implementation for each of the prompt types.
Taken together, the Reflective Lessons were expected to add between 8 and 11 class periods to the sinking and floating Unit. Given that the study was situated in regular classrooms, we did not, nor were we able to, control the amount of time spent by either the control or experimental group. Analysis of the videotapes submitted by teachers indicated that, on average, teachers actually spent more time on the Reflective Lessons than even the more liberal 11-lesson estimate provided in the Guide. (See Table 7; Mean Number of Lessons = 13, SD = 2.83.)

We initially worried that the addition of the Reflective Lessons would make for longer units overall for the teachers. These concerns were confirmed; in fact, the actual number of calendar days (including weekends and holidays) spanned by investigations 1–12 and the Reflective Lessons varied greatly between classrooms (Mean = 139.17 days, SD = 55.43). Although these calculations incorporate many days in which class was not held, the extent to which the unit and Reflective Lessons were spread out across multiple months is still remarkable (e.g., Becca’s 242 days vs. Carol’s 87 days).

Measuring Quality of Delivery of Treatment Processes

Comparing the implementation of treatment processes to what was intended is less straightforward than doing so with treatment structures. There was no definitive sequence given to teachers as to how to conduct whole-class discussions;

---

2Because the estimates of timing are based on the videotapes submitted by teachers, we acknowledge that, beyond videotape we know was lost due to poor audio quality, these figures may be an underestimate of the time teachers spent.
rather, the *Guide* contained many suggested teaching strategies, particular questions to ask, and types of feedback to provide. Therefore, in this section, we do not present an “ideal” implementation profile. Rather, we discuss the critical aspects as described in the *Guide*. Results are presented in the form of proportions because each teacher used different amounts of time to conduct whole-class discussions; to account for this discrepancy, the number of minutes (N) is provided in each analysis. In some cases, the N for a particular teacher varies slightly between codes because some minutes for some codes were dropped from the analysis due to inaudible statements by the teacher and/or students.

(5) **Eliciting student conceptions.** The *Guide* emphasized that, in order to use the reflective lessons effectively, teachers needed to collect and organize student ideas into coherent groups. We classified their strategies for doing so as holding class presentations, asking students to vote to indicate their predictions or reasoning, or simply to ask students to share their ideas (collecting). Figure 4 shows the frequency with which each teacher used these strategies within whole-class discussions.

By far teachers spent most of their time collecting students’ ideas among the four elicitation strategies. This allocation of time to collecting ideas might be expected, because the code served as a kind of “grab bag” for the different ways that teachers asked students to make their thinking explicit. Examination of class presentations and voting revealed important differences among the teachers’ implementations of the formative-assessment treatment. Only two teachers used class presentations regularly as a strategy to make students’ thinking explicit (Diana = 35.7%, Becca = 27.5%), whereas four teachers asked students to share their ideas through voting, for varying amounts of time (mean = 6.25, Min. = 0%, Max. = 13.7%).

### Table 7

<table>
<thead>
<tr>
<th>Reflective Lesson</th>
<th>Total # of class periods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Guide</td>
<td>2–3</td>
</tr>
<tr>
<td>Aden</td>
<td>3</td>
</tr>
<tr>
<td>Andy</td>
<td>3</td>
</tr>
<tr>
<td>Becca</td>
<td>3</td>
</tr>
<tr>
<td>Carol</td>
<td>6</td>
</tr>
<tr>
<td>Diana</td>
<td>3</td>
</tr>
<tr>
<td>Robert</td>
<td>4</td>
</tr>
<tr>
<td>Mean</td>
<td>3.67</td>
</tr>
</tbody>
</table>

*Note.* “0” indicates that we do not have video data for the teachers on this lesson; it is possible that RL elements were implemented and not taped.
(6) Tracking and clustering student conceptions. In order to keep a visual record of what students were thinking, teachers were also asked to track students’ ideas by displaying them in some manner and then grouping those that had common features. Once these ideas were displayed, the Guide asked teachers to cluster similar ideas together so that students could compare them with each other. In fact, the Guide described clustering as essential to using the Reflective Lessons effectively. Figure 5 shows the frequency with which these strategies were used in the classrooms.

All six teachers usually displayed student ideas in some way by using the whiteboard/chalkboard, an overhead projector, poster paper, sticky notes, or smaller pieces of paper that were attached to the board with magnets or tape. However, once these ideas were displayed, only four of the teachers made an attempt to cluster the ideas together in any way (mean percentage of minutes across teachers = 5%, Min. = 0%, Max. = 13.2%).

(7) Asking students to provide reasons for their explanations. Because an important purpose of the Reflective Lessons was to help students advance in their understanding of density and buoyancy, teachers needed to ask students to provide explanations and follow-up questions to get at the conceptions that lay beneath their predictions, claims, and other statements. Figure 6 presents the frequencies with which teachers used these two strategies during whole-class discussions.
The results shown in Figure 6 indicate that most of the time, teachers were not asking students “Why”-types of questions that were intended to get at underlying reasoning; the majority of time in five of the six classrooms involved no instances of teachers asking their students to provide explanations (mean percentage of minutes across teachers = 65.7%, Min. = 45.5%, Max. = 82.4%). In addition, although all teachers did occasionally ask follow-up
questions, only two teachers—Aden and Diana—did so for more than 10% of the time.

(8) Students argue ideas and evidence. The Guide explicitly states that teachers should encourage argumentation through discussions and debate. An important piece of the idea of “debate” is that students respond to each other’s ideas as opposed to responding only to their teacher’s statements and questions. Figure 7 illustrates how often students were involved in this kind of debate.

For the vast majority of whole-class discussion time in all six classrooms, students were speaking to the teacher, not each other (Figure 7; mean = 82.6%, Min. = 73.1%, Max. = 91.2%). In Becca’s class, students never spoke directly to each other, and in Aden and Andy’s class, this form of argumentation occurred less than 5% of the time. Only in Carol’s class was a somewhat larger portion—20%—of whole-class discussion time spent with students addressing each other.

(9) Students provide evidence for their claims. Students should, according to the Guide, support their claims with empirical evidence—that is, provide data from systematic observations conducted in class that relate to the conceptions underlying their explanations. The Guide describes different qualities of evidence, and possible sources. Nevertheless, we simply looked to see if students, in each classroom, provided any evidence when making statements during whole-class discussions. Table 8 presents the results of this analysis.

![FIGURE 7 Percentages of argumentation levels and total discussion time in minutes by teacher.](image-url)
On average, students supported their claims with evidence about 25% of the time; however, there was considerable variation among teachers ($SD = 9.3$, Min. = 11.8%, Max. = 39.4%).

**Learned Treatment**

To examine the relationship between fidelity of RL implementation and student learning, we focused on student performance on the multiple-choice pre–posttest. Although other measures of student learning were assessed as part of the Romance Project (see Yin et al., this issue), the 38-item multiple-choice achievement test was the only achievement assessment administered before and after the treatment, and as such was the only assessment that could provide measures of what students actually learned during the course of the treatment. First, we determined whether or not students’ average achievement differed by teacher on the pretest. A one-way ANOVA indicated, as expected given the different demographic profiles of the participating schools, a significant mean difference between the six groups at the beginning of the study ($F_{(5,139)} = 2.99$, $p = .014$). Tukey’s HSD indicated that the significant difference observed was only between Becca and Robert’s students; however, this difference was not surprising given that the mean score of Becca’s students was the lowest observed and Robert’s the highest among the six classes (see Table 9).

Next, we focused on differences between students in classes at the end of the study. We used an analysis of covariance, controlling for pretest. Results indicated a significant difference in the posttest scores ($F_{(5,120)} = 8.72$, $p = .000$). Post-hoc comparisons indicated significant mean differences between Carol (highest adjusted mean score) and all of the other teachers, but Andy ($p = .057$). Robert’s students (lowest adjusted mean score) were significantly lower than Andy’s, Carol’s, and Diana’s students, but not to Aden’s and Becca’s students.
Comparing Enacted and Learned Curricula

In this section we take the final step in connecting the Intended, Enacted, and Learned Treatment in our examination of the relationship between implementation fidelity and student achievement. Because we did not have videotapes for all lessons for all teachers, there is little validity in making comparisons between enacted treatment structures and student learning. However, because analyses of treatment processes were based on proportions and not total data, we were able to develop a ranking for each of the critical processes based on the extent to which each teacher’s enacted treatment aligned with the intended treatment as defined by the Guide.

The rankings for the Quality of Delivery are based on teachers’ congruence with the Treatment Processes that guided our analysis; meaning, the higher the percentage of time those processes were implemented, the higher the teacher’s ranking. For the first aspect, eliciting student conceptions, no ranking was given because none of the three strategies coded were viewed as being more critical than the others. The second aspect, tracking and clustering students’ ideas, did have a critical aspect identified by the Guide; that of clustering students’ ideas. For the third aspect, asking students to provide reasons for their explanations, we viewed both asking for explanations and asking follow-up questions to be critical; therefore, the teachers who had higher proportions of both of these codes received higher rankings. The fourth aspect, encouraging argumentation, was simpler; teachers whose students talked to each other more were rated more highly. Finally, students provide evidence for their claims also translated readily into a ranking; teachers whose students cited evidence more received higher rankings.

These results and average teacher rankings are shown in Table 10 alongside the posttest–pretest change scores. The means for the enacted treatment rankings and learned change scores were then used to produce overall rankings for the
teachers on the enacted and learned treatments. The relationship between each teacher’s learned and enacted treatment ranking is shown in Figure 8.

Spearman’s rank-order correlation indicated a relationship between the rankings of the enacted and learned treatments; however, the correlation,
although fairly high in magnitude, was not significant, likely due to the small sample size ($\rho = 0.714, p > .05$). The comparison between the rankings does reveal three groupings of teachers; first Carol, who ranked highest on both the enacted and learned treatment; second, Aden, Diana, and Andy, who had intermediate ranks on both measures; and finally Robert and Becca, whose rankings were lowest for both enacted and learned treatment. This relationship, although by no means causal, at least indicates that there was some similarity between the ranking of teachers’ enacted quality of delivery and learned treatment.

DISCUSSION

The goal of this article was to provide information about the quality of implementation of the formative-assessment treatment, the embedded assessments, with the purpose of gaining some understanding about the results reported by Yin et al. (this issue)—large teacher effects and no treatment effect. We argued that in the absence of information about implementation, it would be difficult to determine whether Yin et al.’s results could be attributed to a poor conceptualization of formative assessments or to inadequate implementation.

Results indicated that adherence to the treatment structure varied by type of embedded assessment. Higher levels of adherence to the structure were observed in Reflective Lessons Type I than Type II. It seems that both the Guide and the training emphasized Type I more than Type II Reflective Lessons. It might be that the importance of the information gained about students’ level of understanding through concept maps was not emphasized enough in the Guide, in the training, or by the researchers. Teachers devoted much more time to the discussion of Reflective Lessons Type I.

Teachers’ quality of delivery departed even more from the envisioned implementation than from the treatment structures. Some critical aspects were implemented across teachers (e.g., whole class discussions), whereas some others (e.g., clustering students’ conceptions or asking for students’ explanations) were almost totally ignored by most. Whole classroom conversations and collecting information from students, alone, are not what makes for a good implementation. Clustering students’ conceptions and asking for explanations, elaborations, and supporting evidence are the most fundamental characteristics and intentions of the embedded assessments.

The quality of delivery seemed to be more consistent with teachers’ rankings on the change scores (posttest–pretest achievement) as opposed to the adherence rankings. This result supports the contention that simply giving students the embedded assessments in sequence may not be enough to help students learn; the quality of delivery of the critical teaching strategies is an essential element in helping students learn.
Although we cannot conclude that the variation among the teachers’ implementations of the treatment led to the differences in student learning reported by Yin et al. (this issue), the results at least suggest a correlation between the consistency of treatment enactment with the project’s intention and student learning.

Implementing quality formal formative assessment requires a careful consideration of design and practical issues not discussed with enough depth in the literature. Formative assessment tasks should be designed not only to make students’ conceptions explicit but also to do it in an efficient and effective way so the embedded assessments do not take a lot of time to be administered and focus attention on the critical issues to which the teacher and the students need to pay attention. How much do we know about the types of assessment task that best reflect students’ conceptions? Few attempts have been made to learn about assessment tasks in the context of classroom assessment (but see Furtak & Ruiz-Primo, 2007). We believe that the data collected in this study indicate the need to develop embedded assessments that are easy not only to implement, but also to gather student’s conceptions that clearly reveal what steps need to be taken to close the formative assessment loop (“how to get there,” the third step in the three-step cycle proposed by Ramaprasad in 1983).

The results of this study suggest that the ADT did not place a clear enough emphasis on what it was considered the critical aspects of the treatment. It is possible that the complex structure of the Reflective Lessons made them difficult to learn to teach in one week, and as a result, the critical aspects that were implemented more—the treatment structures—may have been easier for the teachers to learn than the treatment processes.

The study also raises important questions about the feasibility of formative assessment in general. For instance, embedded formative assessments may be too restrictive and time-consuming for teachers. In addition, some of the most critical aspects of implementation, for example, pushing students to support their claims with reasons and evidence, or encouraging students to argue with each other, are difficult for any teacher, especially those with limited teaching experience and weak backgrounds in science. One possible solution may be to avoid focusing on the structure of formative assessments and instead work with teachers to improve their ability to lead whole-class discussions that truly engage learners in sharing and arguing their ideas.

In hindsight, we believe that we should have put less effort into presenting teachers with many possible teaching strategies, and more effort into identifying what we believed were the most important strategies to help students learn. Furthermore, a model to guide teachers’ quality of delivery may have helped explicate our instructions to “Cluster student ideas” and “argue and debate,” although the effectiveness of this model would be entirely dependent on teachers’ willingness to use it—as we found with the materials we already provided them.
A possible model might be for teachers to gather student ideas, display them, and then cluster and discuss them, seeking supporting evidence.

Perhaps the most important lesson we learned in the process of completing the Romance Project is that, despite our best intentions to design a treatment and instruments to measure it, the actions of the project’s ideas were only as good as our ability to help teachers enact them in the classroom.

ACKNOWLEDGMENTS

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

REFERENCES


