

**A Multi-Method and Multi-Source Approach for
Studying Fidelity of Implementation**

CSE Report 677

Maria Araceli Ruiz-Primo
SEAL, Stanford University/CRESST

February 2006

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE),
Graduate School of Education & Information Studies
University of California, Los Angeles
GSE&IS Building, Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

Program Two: Tools & Support for Cognitively Oriented Tests. Project 2.2: Classroom and Teachers' Assessment.
Richard J. Shavelson & Maria Araceli Ruiz-Primo

Copyright © 2006 The Regents of the University of California

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Institute on Student Achievement, Curriculum, and Assessment, the Office of Educational Research and Improvement, or the U.S. Department of Education.

A MULTI-METHOD AND MULTI-SOURCE APPROACH FOR STUDYING FIDELITY OF IMPLEMENTATION

Maria Araceli Ruiz-Primo
SEAL, Stanford University/CRESST

Abstract

Even the best program in education will fail to have the intended impact if its essential elements are not implemented properly. Degree of implementation is, then, critical to draw valid conclusions on program outcomes (e.g., Scheirer & Rezmovic, 1983). Especially important is the information on the fidelity with which a program is implemented. *Fidelity of Implementation* (FOI) has been defined as the determination of how close the program is implemented according to its original design or as intended (e.g., Dobson & Shaw, 1988; Dusenbury, Brannigan, Falco, & Hanse, 2003; Witt & Elliot, 1985).¹ Unfortunately, empirical evidence on the effect of FOI on program success is limited. Many evaluation studies do not collect data on FOI and even fewer examine its impact on program outcomes (Dane & Schenider, 1998; Dusenbury et al., 2003; Lillehoj, Griffin, Spoth, 2004). Furthermore, studies on FOI differ considerably on their approaches (Dane & Schenider, 1998; Dusenbury et al., 2003; Huntley, 2004; Lillehoj, Griffin, Spoth, 2004); there is no set of methods and procedures that is universally known and used as standard procedure in the study of FOI. Whereas the characteristics of each program determine what has to be measured during implementation, there are some commonalities across types of programs and, therefore, some general strategies that can be developed.

This paper addresses FOI at three levels: general, conceptual, and applied. The first section provides a short review of literature on the main issues of FOI. The second section proposes a conceptual approach for studying FOI in the context of inquiry-based science curricula. The third section describes a series of studies, currently in progress, in which this conceptual approach is being used.

¹ Some authors also refer to fidelity of implementation as *integrity verification* (Dane & Schneider, 1998) or *treatment integrity* (Dobson & Shaw, 1988; Gresham, 1989; Waltz, Addis, Koerner, & Jacobson, 1993).

Fidelity of Implementation: General Issues

Defining Fidelity of Implementation

No consensus exists on what exactly constitutes FOI (e.g., Fullan & Pomfret, 1977; Scheirer & Rezmovic, 1983). This inconsistency makes it even more difficult to interpret studies that focus on FOI and how it relates to outcomes (Dane & Schneider, 1998; Lillehoj, Griffin, Spoth, 2004). A review by Dane & Schneider (1998) helps to bring consensus on some aspects for measuring FOI. Based on definitions found on diverse evaluation studies, these authors presented five aspects of FOI that have been measured across studies: (a) *adherence* – extent to which specified program components are delivered as the program prescribes; (b) *exposure* – amount of program content received by participants (i.e., number or length of sessions or frequency with which program techniques are implemented); (c) *quality of program delivery* – extent to which providers approach a theoretical ideal in terms of delivering program content and processes²; (d) *participant responsiveness* – extent to which participants are engaged; and (e) *program differentiation* – uniqueness of the features of the program or treatment components that can be reliably distinguished from others.³ In a later review, Dusenbury et al. (2003) revised these definitions, but their essential meaning did not change. Dane and Schneider (1998) recommend to measure all five aspects in order to provide a comprehensive picture of the fidelity of the program.

Studying FOI helps to understand how the degree of program implementation can affect the achievement of the goals and, more importantly, how implementation can be improved when the program needs to be disseminated or scaled up. It seems, then, that if data on fidelity were not collected, it would be difficult to determine whether non-significant results are due to poor program conceptualization or to inadequate program enactment (Bauman, Stein, & Ireys, 1991; Boruch & Gomez, 1977; Dane & Schneider, 1998; Dobson & Shaw, 1988; Gresham, 1989; Moncher & Prinz, 1991). Unfortunately, while most of the reviews have been conducted in the area of health prevention and behavior therapy (Dobson & Shaw, 1988; Dane & Schneider, 1998), few are known in education (Berman & MacLaughlin, 1976; Fullan & Pomfret, 1977; Snyder, Bolin, Zumwalt, 1992).

² Waltz, Addis, Koerner, & Jacobson (1993) referred to this aspect as *competence*.

³ Dobson and Shaw (1988) used the term *treatment differentiability*.

There are some interrelated reasons to collect information on FOI. First, differential patterns of implementation can be linked more easily to differential program effectiveness (Blakely et al., 1987; Boruch & Gomez, 1977). Second, FOI provides information to developers on the omission of the implementation of critical characteristics of the program or the implementation of inappropriate techniques. FOI, then, helps to document deviations from the intended program (Mowbray, Holter, Teague & Bybee, 2003) and to identify some of the most problematic aspects in a program's implementation (Fullan & Pomfret, 1977). Third, FOI provides information that can help fine-tune the program, training manuals, training facilitators, and program supervisors. Fourth, lack of standardization within and between program providers will inflate error variance and decrease power (Boruch & Gomez, 1977; Moncher & Pinz, 1991). Fifth, FOI information helps evaluators and researchers understand why some programs fail to become established or transferred, (Bauman, Stein, & Ireys, 1991; Fullan & Pomfret, 1977).

It is clear that "successful" projects have not been easy to export across schools or districts or results cannot be consistently replicated (Bauman, Stein, & Ireys, 1991; Berman & MacLaughlin, 1976; Dane & Schenider, 1998). Information on fidelity is therefore necessary for its implications to validity (interpretation of program outcomes) and replication. Furthermore, it helps to understand the nature of the program (treatment or curriculum). We do not know what is changed during the implementation of a program unless we attempt to conceptualize and measure it directly (Bauman, Stein, & Ireys, 1991; Fullan & Pomfret, 1977). FOI, then, has implications for internal, external, and construct validity (Moncher & Prinz, 1991).

Specific purposes of FOI can be linked to the stage of development of the program at hand.⁴ At the experimental phase, conducting FOI studies can include purposes such as: (a) identifying variations in its implementation, (b) identifying conditions under which the implementation of the program at hand is likely to succeed (including problems likely to be encountered and strategies available for their resolution), (c) determining the capabilities needed for proper implementation, (d) determining whether the implementation of the program is warranted for assessing its effectiveness in achieving its goals, or (e) determining which components can be associated with its effectiveness. At the prototype level, FOI studies can include purposes such as identifying differential implementation between sites to define the issues that are

⁴ The development of a program can be conceived in three stages: planned, experimental and prototype (Ruiz-Primo, 1994; Ruiz-Primo, Shavelson & Baxter, 1995).

critical to scaling-up the program. It can be argued, then, that studying FOI is more relevant to formative evaluation, when modifications to the curriculum or program can still increase feasibility. The issue of determining what works best for whom is not an idle pursuit. However, in an era of experimental trials, FOI can also be of relevance. Experimental or quasi-experimental studies in which the treatment involves some kind of innovative instructional practice have often focused on comparing experimental and control groups assuming a dichotomous categorization between treatment and control groups. This perspective assumes that all students in the intervention group receive comparable treatment. However, in reality, large variations may likely characterize the implementation of the treatment (Harachi, Abbott, Catalano, and Fleming, 1999).

Still, some have argued that fidelity is of marginal importance, since adaptation is necessary for a program to be successful (Berman & MacLaughlin, 1976; Hord, Rutherford, Huling-Austin, & Hall, 1987). However, it can be argued that fidelity of a program must be maintained at the level of the program's mechanism of operation; that is, the causal mechanisms (related to the outcome) of the program must be preserved (Baum, Stein, & Ireys, 1991; Boruch & Gomez, 1977).

Factors Related to FOI

Several characteristics of a program affect the fidelity of its implementation. Figure 1 provides a summary of the factors affecting FOI organized according to two dimensions: program characteristics and setting context.

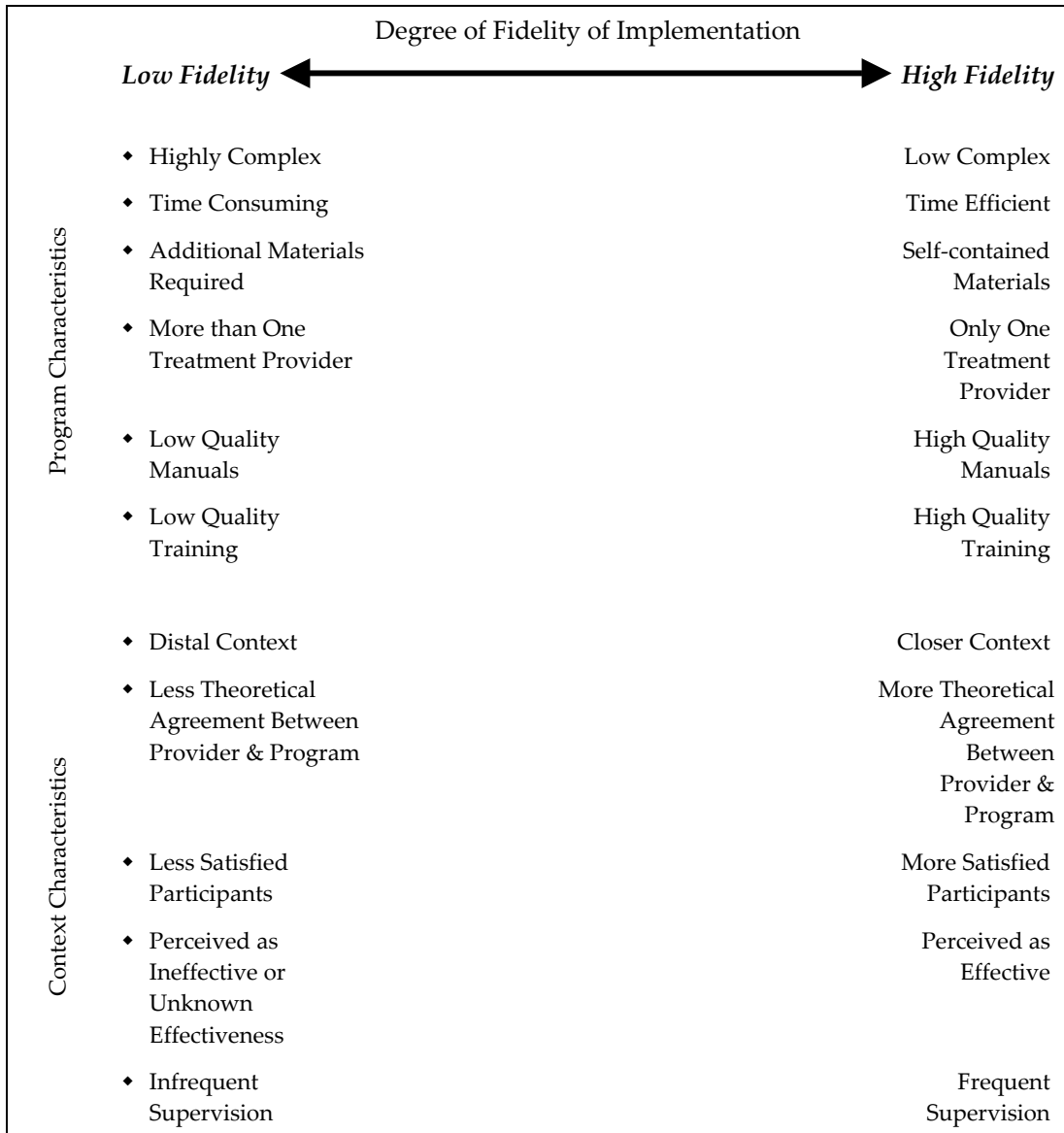


Figure 1. Factors that can affect the fidelity of its implementation according to the Program, Setting, and Implementation Characteristics.

Complexity of the program (or treatment) has been cited consistently as a factor affecting fidelity. The more complex the treatment, the lower the fidelity (Bauman, Stein, & Ireys, 1991; Gresham, 1989; Witt & Elliot, 1985). Complexity can refer to the number of interrelated program components, to the steps involved, to the precision or coordination requirements, or the difficulty in grasping what makes the program work;

what makes it effective.⁵ *Time required to implement the program* is a factor that interacts with the complexity of the program. The more complex the program, the more time is required for its implementation. The longer a program takes the less likely that is implemented with fidelity (Gresham, 1989).

Materials and resources required for the program is another factor to consider. Programs that require additional materials and resources are likely to be implemented with poorer fidelity than planned (Gresham, 1989). *Number of providers* also affects fidelity (Gresham, 1989). Programs requiring more than one provider maybe implemented with less fidelity than programs requiring one provider. *Implementation manuals or guides* have proved to enhance fidelity. However, for achieving fidelity, quality matters. Manuals and guides should provide explicit guidelines for techniques and strategies that comprise acceptable implementation of a given program or approach. Appropriate level of specificity matters – too molecular is overwhelming but too global is inadequate (Moncher & Prinz, 1991). Manuals and guides should be both prescribed and proscribed since it is equally important to know about techniques and strategies that are inconsistent with the program (treatment) approach (Dobson & Shaw, 1988). They also should provide criteria for evaluation of competency. That is, high quality manuals should facilitate the decision about when a program provider is trained to the level that is representative of the program approach. (Dobson & Shaw, 1988; Moncher & Prinz, 1991). Clearly, manuals alone are insufficient to ensure fidelity. *Training* is fundamental to support fidelity by marking boundaries for the delivery of the program (Moncher & Prinz, 1991). If the training does not make clear what exactly makes a program successful, the fidelity of its implementation will be reduced (Gresham, 1989; Witt & Elliot, 1985).

Not all the factors that affect fidelity are directly related or determined by the characteristics of the program. There are other factors that are related to the sites and its context (Bauman, Stein, and Ireys, 1991). *Sites context* with characteristics closer to those envisioned by the program developer are likely to have higher fidelity than those with characteristics not originally considered in the design of the program (Bauman, Stein, & Ireys, 1991). Characteristics such as staff experience or commitment are known to affect FOI (Bauman, Stein, & Ireys, 1991; Moncger & Prinz, 1991). However, others are forgotten. *Theoretical agreement* is one of them. Programs will be implemented with

⁵ Complexity can involve concepts, skills or content that are new to the providers, and therefore be difficult to deliver with fidelity. It can be the case that program developers do not clearly delineate which key components exactly make a program effective and successful.

higher fidelity when there is an agreement between the provider and the program approach (Moncher & Prinz, 1991). *Perceived effectiveness* also affects FOI. Programs that are perceived by providers to be effective may be implemented with greater fidelity than those perceived to be ineffective or in which effectiveness is unknown (Gresham, 1989). Similar conclusions are reached with respect to the perceived effectiveness of the program by participants (Witt & Elliot, 1985). In prevention or health programs in educational settings, *participants' acceptability* or *satisfaction* affects the fidelity of the program. Participants follow instructions or conduct the necessary activities better when they like the program than when they do not (Witt & Elliot, 1985). *Supervision of providers* is one factor considered differently across sites. Frequent supervision increases fidelity, especially if feedback is provided (Moncher & Prinz, 1991).

Relation of FOI and Outcomes

Few research reviews have been published that provide a good picture of where FOI studies stand in relation to outcomes. Berman & MacLaughlin (1976) and Scheirer and Rezmovic (1983) are pioneers.⁶ Recent reviews (Dane & Schneider, 1998; Dusenbury et al., 2003; Gresham, Gansle, Noell, Cohen, & Roseblum, 1993; Lillehoj, Griffin, & Spoth, 2004) provide some insight into this relation.⁷ Gresham et al., (1993) reviewed 181 behavioral intervention studies published between 1980 and 1990. From those, only 14.4% measured and reported fidelity. They found moderate but significant correlations between degrees of fidelity (percent) and level of treatment outcomes (e.g., effect size). They concluded that higher fidelity was associated with larger effect sizes. Dane and Schneider (1998) reviewed 162 evaluation studies of prevention programs (e.g., drugs). They focused on the extent to which fidelity was promoted (e.g., manuals and training are provided to program providers) and how the effect of fidelity on outcomes was approached. These authors found that only 24% of the papers documented some sort of FOI. Only 13 studies analyzed outcomes as a function of one or more aspects of fidelity. In general, they concluded that the evidence on the relation between FOI and program outcomes is inconsistent. Results vary depending on both the aspects of FOI considered for evaluating this relation and how these aspects are measured. This was the case of all

⁶ The review of Fullan and Pomfret (1977) did not focus on examining the effect of quality of implementation with student outcomes, although some information is provided for some of the studies.

⁷ Moncher and Prinz (1991) reviewed 359 studies to examine the trends in assessing fidelity across three time periods (1980-82, 1983-85, and 1986-88). Unfortunately, they did not relate the effect of fidelity on outcomes. Their analyses focused on three areas: promotion and verification of fidelity, source, sampling, and utilization of fidelity procedures, and training of program providers. They found that 55 percent of the studies reviewed ignored the issue of treatment fidelity.

aspects except for one—exposure. Dane and Schneider (1998) concluded that exposure (dosage) is positively related to effectiveness in prevention programs; programs appear to be less efficacious for subjects who received a lesser proportion of the program.⁸ Similar results were concluded in an implementation study of an individualized mathematics early learning program. Leinhardt (1974, cited by Fullan & Pomfret, 1977) concluded that 35% of the variance observed in achievement was explained by the degree of implementation.

Blakely et al. (1987) studied the differential effectiveness of fidelity versus adaptation in seven national education and criminal justice projects. They found that programs adopted with high fidelity were more effective than those which had been modified or adapted with some qualifications: Local additions (addition of something new to the program components) increased its effectiveness; however, local modifications (change of existing program components) were less effective or unrelated to effectiveness.⁹

In sum, it seems that there is some empirical evidence that the effectiveness of a program can be associated with the fidelity with which it is implemented, regardless of how program outcomes are measured.

Measuring FOI

Defining program components. The first challenge in assessing program implementation is defining what we are looking for (Gresham, 1989; Hall & Loucks, 1977; Moncher & Prinz, 1991; Mowbrey et al., 2003). That is, *will you know it when you see it?*¹⁰ Answering this question goes beyond defining the goals or describing general requirements for its implementation. To measure fidelity of a program's implementation, first we need to define exactly what the program is. "Nothing blocks communication, inhibits evaluation, hampers staff development, and thwarts improvement more than a program that is not clearly defined" (Crandall & Loucks, cited in Alliance Access, p. 3; see Footnote 9). Clear specification of what the program entails is necessary to ensure that the active ingredients of the program are being delivered (Moncher & Prinz, 1991). Ambiguity decreases fidelity and transferability. It is important, then, to know *why* a program works – the essential components of the

⁸ Dusenbury et al. (2003) reviewed many of the papers reviewed by Dane and Schneider (1998), therefore, the conclusions are similar.

⁹ If a program component is omitted altogether, it was viewed as an unacceptable variation or lack of fidelity.

¹⁰ Adapted from Alliance Access, Vol 3 # 3 without author.

program and the process through which the program achieves its effects (Bauman, Stein, & Ireys, 1991).

Hord, Rutherford, Huling-Austin and Hall (1987) conceptualized social programs as consisting of a finite number of components. If the program is defined in parts or components, it is easier to know how the different components are being implemented by the program providers (e.g., teachers). *Component* refers to major operational features or parts of the program (Hord et al., 1987). When these components are determined to be essential in the program, they are designated as *critical* or *crucial* (Gresham, 1989; Hord et al., 1987) or as having a greater weight (Gresham, 1989). Those which are not critical are named *related*. They are recommended components, but not critical ones (Hord et al., 1987). Program fidelity, then, can be defined as the number or proportion of finite program critical components that are implemented (Gresham, 1989; Hall & Loucks, 1977).

How should we determine critical and related components? For some, the designation of program components as either critical or related should be done by the program developers, users, and evaluators. Preferably, this process should be through the consensus of all these parties. Providing the developer's perspective only leads to a conservative model of program components (Blakely et al., 1987). In the view of others, components should be explicitly recognized by the providers in order to elicit valid information (Leithwood & Montgomery, 1980). In their review of FOI studies, Mowbray et al., (2003) refer to this method as *qualitative research* – opinions of users and advocates regarding the value of the components. Others prefer an *empirical method*. For example, Cook, Leviton, and Shadish (1985) consider that the critical components should be those that are necessary for effectiveness of the program. Proven efficacy and functional analysis (Haynes & O'Brien, 1990) are necessary. The latter one helps to distinguish which program components are necessary (if they are not implemented as planned, the program outcomes will not be reached) from those that are not, but are related to achieving the outcomes (see Haynes & O'Brien, 1990). Mowbray et al. (2003) used a third category, *gathering expert opinions*, which entails collecting information from surveys of experts and/or literature reviews.

Another issue to consider in component specification is the level of specificity required (Gresham, 1989). What is best, global, intermediate, or molecular levels of specification? On the one hand it makes sense to be specific, but on the other, it may be overwhelming and not very practical. Intermediate level seems to be the way to go, but how can we define intermediate? Gresham does not provide an approach to solve this

problem. At any level, components need to be defined in a way that can be measured and that is parallel to the program developers' description.

Another consideration in assessing fidelity is the definition of *degree of deviation*. How far can implementers deviate and still achieve the goals? Is it possible to deviate from the prescribed plan and achieve the program goals? In general, there is no empirical evidence to make this decision. There is no question that within each component are *variations* – the different ways in which the program provider can put a component into operation – but, what is the point in which the mutation is drastic? (Hord et al., 1987). The important issue, then, is to determine the degree of variation that can be considered acceptable and how it affects the nature of the critical components. Boundaries must be always specified (Haynes & O'Brien, 1990). Hord et al. (1987) suggested having *configurations* that reflect different patterns of implementation. In order to have configurations of critical components, it is important to define the degree of variation; which ones are minor variations and which ones are major, which ones will be considered acceptable and which ones will not (Bauman, Stein, & Ireys, 1991).

Unfortunately, few studies provide enough information about the program components, variations, or configurations (see also Mowbray et al., 2003). This information could help to define components that need to be considered in FOI, based on the nature of the programs (e.g., prevention, behavioral, or educational). Table 1 presents some examples of program components cited in a few studies.

Table 1. Examples of Components Considered In Educational Programs

Authors	Components Evaluated
Leinhardt (cited in Fullan & Pomfret, 1977)	Context Allocation of time and space Assignment procedures Classroom management Student independence
Evans and Scheffler (cited in Fullan & Pomfret, 1977)	Placement of pre-post test Curriculum embedded tests Prescription writing Classroom management Student self-management Planning session
Hord et al. (1987)	Materials Teacher behaviors Student activities
Hoolbrook, Gray, Fasse, Camp & Kolodner (2000)	These are only examples. Paper was not clear on how many components were measured: Students use of vocabulary/measurement techniques w/accuracy Students listen/discuss/consider ideas/suggestions of other w/in group Teacher knowledge of the specific science areas Teacher shows flexibility for changing plans when indicated by student needs
Mowbray et al. (2003)	Specification of the length, intensity, and duration; Content, procedures, and activities over the length of the service Roles, qualifications, and activities of staff Inclusion/exclusions characteristics for the target population
Schneider, Krajick and Blumenfeld (2005)	Presentation of science ideas Opportunities for student learning Support to enhance the learning opportunities

Clearly, there is a variety in the program components measured in diverse programs. Most of them, naturally, focus on teachers' behaviors, and few on the students. The variation can be higher when talking about the criteria to measure level of fidelity and how the levels are measured. For example, Schneider, Krajick and Blumenfeld (2005) used eight categories to analyze the instructional events: accuracy of

the scientific ideas, completeness of the scientific ideas, number or time opportunities for student learning, similarity with intended lesson, level of adaptation, level of instructional support to students, appropriateness of instructional support, and sources used for instructional support. Categories were coded using a four-level rating. Only one category was a five-level rating. However, focus of the levels varied from category to category (e.g., from high to none, excellent to poor, maximum to minimal, or scientific to nonscientific). Hoolbrook, Gray, Fasse, Camp & Kolodner (2000) rated fidelity in a five-point Likert scale (i.e., unsatisfactory, needs much improvement, meets expectations, good, and ideal). These authors also provide examples of configurations by defining variations across the five-scale use levels for the same component.

Instruments and sources of information. Figure 2 presents categories of instruments and sources of information used in most of the FOI studies organized according to four dimensions: *extent of judgment* – assessment method is more or less objective (Scheirer & Rezmovic, 1983); *directedness* – extent to which the assessment methods directly captures the implementation; *sensitivity* – degree to which it can detect the actual providers and participants behavior (Scheirer & Rezmovic, 1983); and *alignment to the program* – extent to which the instruments detect program characteristics. Three main categories of sources of information were considered: program provider or implementator (e.g., teachers), participants, and independent observer. The latter category can include program sponsors, program developers, or trained independent observers or researchers. Finally, the figure considers the alignment of the instruments to the program. Some instruments are developed based on the program; some others are adopted from other programs. Therefore, the alignment may vary based on how close the instrument is to the characteristics of the program. Factors to be considered in selecting and designing fidelity instruments are: purpose, subject matter of the observation, amount of behavior to be observed, and quality of the data produced.

Observational techniques represent the most rigorous measurement of FOI according to some researchers (Fullan & Pomfret, 1977). No doubt, most of the FOI studies involve observational techniques. It is clear that some dimensions of program components are more difficult to assess through observation than others. There is a chance that observation taps only mechanical use but does not adequately measure other dimensions of the program components that are more relevant (e.g., degree of understanding the philosophy or strategies and techniques). Still, direct observation methods are, in general, unfeasible if many providers and sites are involved.

Extent of Judgment	Directedness	Sensitivity	Alignment with Program	Source
<i>Low</i> ↑ ↓ <i>High</i>	<i>Indirect</i> ↑ ↓ <i>Direct</i>	<i>Low</i> ↑ ↓ <i>High</i>	<i>Low</i> ← → <i>High</i> Measure of an output - Products Documentary Analysis Rating Scales Direct Observation (rating scales) Direct Observation (checklist) Interviews, Questionnaires, & Logs Direct Observation (notes) Self reports Ethnographic Observation	Independent Observer Independent Observer Independent Observer Independent Observer Independent Observer Provider/Participant Independent Observer Provider Independent Observer

Figure 2. Techniques for measuring FOI, ordered by extent of judgment, directedness, sensitivity and alignment with the program.

Others consider that if the program components are clearly and operationally defined, FOI can focus on the record of occurrence and non-occurrence of the key components (Gresham, 1989). This procedure leads to a straightforward indicator of level of fidelity, the percentage of treatment components implemented by providers over the total implementation of a program or over time (e.g., in a week; see Gresham, 1989).

Fidelity of Implementation: An Approach to Studying Science Curricula Implementation

In this section, I focus on describing an approach to measure FOI of science curricula. The approach considers the nature of the program type and it draws on both literature on FOI (discussed above) and literature on science education and science inquiry.

The approach involves three elements: types of curriculum, curriculum dimensions, and the aspects of fidelity. The types of curriculum are (Schmidt et al., 1996): (a) The *intended curriculum*, that refers to the content, pedagogy, and structure expressed in the instructional materials that reflect the developers' theory of knowledge

and skill acquisition; (b) the *enacted curriculum* that refers to the way teachers deliver the instructional materials; and (c) the *achieved curriculum*, what students experience and integrate in their existing knowledge and skill structure.

Curriculum is thought as having three major dimensions: content, process, and outcomes. However, for FOI purposes, I propose four major dimensions (Leithwood & Montgomey, 1980; Madaus & Kellaghan, 1992; Mortimer & Scott, 2000): (a) *theoretical stand* refers to a system of implicit and explicit beliefs and assumptions used as the basis for deciding the characteristics that the curriculum have; (b) *curriculum materials* that consider content, activities that have been put together in an specific sequence and that can take different forms of documentation; (c) *instructional transactions* between the teacher and the students that involve the teacher interventions enacting the curriculum and directing it towards making scientific knowledge available to students, and (d) *outcomes* that reflect the intended goals for students.

Finally, the approach considers the five aspects involved in measuring FOI described previously by Dane & Schneider (1998) and Dusenbury et al. (2003): (a) *adherence* – extent to which specified curriculum components are delivered as the curriculum prescribes them; (b) *exposure* – coverage of the curriculum (e.g., investigations implemented, length of time on each investigation, concepts taught); (3) *quality of curriculum enactment* – consistency between the pedagogical ideals (skills, techniques, or methods prescribed) outlined in the curriculum and how teachers are enacting the curriculum in the classroom; (d) *student responsiveness* – student involvement in the curricular activities from discussion to small groups; and (e) *curriculum differentiation* – identifying unique curriculum components that differentiate them from others.¹¹

In what follows, I describe the approach using the second element, curriculum dimensions, as a guide. First, I describe the curriculum dimension at hand from the intended perspective. Then, from the enacted perspective, I propose some critical components to be considered in FOI studies as well as some criteria to consider for determining the level of fidelity. Finally, I link the component and criteria to aspects of FOI. Figure 3 presents the three elements and the different critical components to consider in FOI studies. Two issues are important to note. Firstly, the intention is not

¹¹ Oaks, Gamoran, & Page (1992) use curriculum differentiation to refer to the characteristics of the curriculum that allow that different knowledge is available to different groups of students. From the program evaluation perspective, this meaning may or may not be a factor to consider in differentiation.

that all of the components and the criteria suggested in this approach be measured on every FOI study of science curricula. The intention is to guide the attention of researchers or evaluators towards relevant issues about the implementation of a science curriculum that may affect its effectiveness. Secondly, it is important to consider that FOI measurement instruments can tap more than one component. Therefore, characteristics of instruments design are essential for making a FOI study more efficient.

Theoretical stand. This dimension focuses on the *implicit* and *explicit* beliefs and assumptions for deciding what will constitute the curriculum: (a) What we want students to *know* and what they need to *do* to *know* it (Duschl, 2000; Leithwood & Montgomey, 1980); (b) What is the planned and the intended context of implementation (Leithwood & Montgomey, 1980; Madaus & Kellaghan, 1992); and (c) What teachers' beliefs and values are important to properly implement the curriculum at hand (Kennedy, 2004).

Clearly, site context and teachers' beliefs and values cannot be directly linked to the five aspects of FOI. However, if measured, they can help to make the connection between context characteristics, levels of fidelity on the different aspects of FOI, and effectiveness of the curriculum. I suggest two fidelity criteria linked to theoretical stand: *suitability* can focus on the matching level between the characteristics of the site and the characteristics that are appropriate for the implementation of the curriculum. The second criteria could be *compatibility* between teachers' beliefs and values and those required by a curriculum. Based on the research conducted by Kennedy (2004), it is clear that teachers' "standing beliefs and values," as she named them, are displayed in specific situations and involve specific actions. These beliefs and values include ideas that tend to be deeply held and that are relatively less malleable. Kennedy (2004) mentions the following beliefs and values to consider: general theories of student learning, theories of student motivation, beliefs about the teachers' role and responsibilities, and beliefs about the nature of subject matter and what is important to know about it.

<i>Intended</i>	<i>Enacted</i>					<i>Achieved</i>		
<i>Curriculum Dimensions</i>	Curriculum Critical Components	Focus of Levels of Fidelity	Aspects of Fidelity of Implementation			Instruments		
			Adherence	Exposure	Quality of Enactment		Student Responsiveness	Curriculum Differentiation
I. Theoretical Stand A system of implicit and explicit beliefs or assumptions that guide the curriculum	Site Context <ul style="list-style-type: none"> Physical resources Teachers' characteristics Students' characteristics School dispositions 	<ul style="list-style-type: none"> Suitability 						
	Teachers' beliefs and values [about] <ul style="list-style-type: none"> Content and what is important to teach and learn How students learn How to support students How to identify student's needs 	<ul style="list-style-type: none"> Compatibility 						
II. Curriculum Materials Content and activities that have been put together in a specific sequence and that can take different forms of documentation	Content and Activities and Their Sequence <ul style="list-style-type: none"> Critical content/ideas Critical activities Critical sequence 	<ul style="list-style-type: none"> Completeness Similarity Quality of Adaptations 	✓	✓				
III. Instructional Transactions Teacher interventions directed towards making scientific knowledge available to students through interacting with them	Developing & Using Scientific Knowledge <ul style="list-style-type: none"> Science Content - Ideas Scientific Approach to Inquiry Science as a Social Process 	<ul style="list-style-type: none"> Accuracy Thoroughness 	✓	✓	✓			
	Providing Learning Opportunities <ul style="list-style-type: none"> Social and physical environment Strategies: <ul style="list-style-type: none"> Group Work Discussion Questioning Instructional Activities <ul style="list-style-type: none"> Conducting Investigations, collecting data, identifying patterns, formulating explanations, evaluating quality of explanations Transferring tasks 	<ul style="list-style-type: none"> Similarity Quality of adaptations 	✓	✓	✓	✓		
	Supporting Student Learning <ul style="list-style-type: none"> Guiding students learning <ul style="list-style-type: none"> Conveying unit/lesson purpose Bridging ideas within/ between lessons Scaffolding (modeling, feedback, etc) Taking account of students' ideas Sharing students' ideas/understanding <ul style="list-style-type: none"> Shared individual ideas w/whole class Shared group findings w/ whole class Jointly discuss an idea w/ a student Checking of student understanding <ul style="list-style-type: none"> Elaboration/clarification/representation of student's ideas Check individual student understanding Check classroom consensus 	<ul style="list-style-type: none"> Thoroughness Appropriateness 	✓	✓	✓	✓		
V. Outcomes Goals for students. Knowledge and skills that students are expected to achieve								Assessment tools tapping what students know and can do

Figure 3. FOI approach for science curricula and three linked elements, type of curriculum, curriculum dimensions, and aspects of FOI

Curriculum materials. This dimension focuses on the content, instructional materials, strategies, and learning experiences put together in a specific sequence and expressed in different forms of documentation such as lesson plans, guides, supplementary reading materials, workbooks, equipment, and/or audiovisuals (Leithwood & Montgomey, 1980; Madaus & Kellaghan, 1992). It is expected that this set of aspects is developed around relevant phenomena and provide the learning conditions that help students to learn science, or as Duschl (2000) puts it, to help students to become members of an epistemic community.

As previously mentioned, curricula as any other program, has critical and related components. For FOI purposes, the identification of the critical concepts or ideas, activities, and the sequence in which they are to be enacted is important. Three criteria can be used to measure the level of fidelity of the curriculum materials component: *completeness* – all the appropriate critical topics and activities (including the implementation of embedded assessments if necessary) are enacted (Penz et al., 1990; Schneider, Krajcik, & Blumenfeld, 2005); *similarity* – closeness of the intended lesson with an enacted lesson based on sequence and major/minor changes (Schneider, Krajcik, & Blumenfeld, 2005); and *quality of adaptations* – adaptations consistent with learning goals and appropriate for students (Schneider, Krajcik, & Blumenfeld, 2005). Completeness can be clearly linked to exposure or dosage (Dane & Schneider, 1998; Dasenbury et al., 2003). Similarity and quality of adaptations tap adherence (Waltz et al., 1993).

Instructional transactions. This dimension involves teacher interactions with students with the purpose of enacting the curriculum materials.¹² It involves three components (Mortimer & Scott, 2000; Schneider, Krajcik, & Blumenfeld; 2005): developing and using scientific knowledge, providing learning opportunities, and supporting student learning.¹³

The first component, *developing scientific knowledge*, refers to the teacher interventions oriented to making scientific ideas available to students. These ideas include the three aspects of public understanding of the nature of science proposed for Driver et al. (1996) and developed by others (Duschl; 2000; 2003; Mortimer & Scott, 2000): (a) science content or ideas – teacher interventions focusing on addressing new and key scientific ideas; (b) scientific approach to inquiry – teacher interventions

¹² Mortimer & Scott (2000) refer to them also as pedagogical interventions.

¹³ During the enactment of the curriculum materials, these three aspects of instructional transactions blend together making the teaching narrative continuous (Mortimer & Scott, 2000).

focusing on aspects of the nature of the scientific knowledge; and (c) science as a social process – teacher interventions involved in students’ scientific oral, written, or pictorial communications. Science content and ideas is what Duschl (2003) refers to as conceptual structures that involve deep understanding of concepts and principles as parts of larger scientific conceptual schemes. Scientific inquiry requires knowledge integration of those concepts and principles that allow students to use that knowledge in an effective manner in appropriate situations. Scientific approach to inquiry is what Duschl (2003) named epistemic frameworks that emphasize not only the abilities involved in the processes of science (e.g., observing, hypothesizing, and experimenting, using evidence, logic, and knowledge to construct explanations), but also the development of the criteria to make judgments about the products of inquiry (e.g., explanations or any other scientific information). Finally, science as a social process refers to the frameworks involved in students’ scientific communications needed while engaging in scientific inquiry. It involves the syntactic and semantic structures of scientific knowledge claims, its accurate presentation and representation, and the use of diverse forms of discourse and argumentation. Criteria for assessing levels of fidelity can include the accuracy and thoroughness of the scientific ideas presented: *accuracy* – content or ideas involved in teacher transactions are consistent with current scientific ideas (scientific vs. nonscientific); and *thoroughness* – all the appropriate science ideas tapped by the curriculum are addressed.

The second component, *providing learning opportunities*, considers the enactment of instructional activities and strategies that are expected to move students forward in their learning. In the context of science education and science inquiry, these learning opportunities should be somehow similar across science curricula. These learning opportunities portray a picture of a science classroom aligned to what it is described in the National Science Education Standards (NRC, 1996), and in the Inquiry and the National Science Education Standards (NRC, 2001). This portrait involves students working in small groups conducting investigations around scientifically-oriented questions, collecting data or having meaningful discussions around procedures, data patterns and evidence-based explanations. Students are expected to be involved in classroom discussions facilitated by the teacher around their conceptions on scientific ideas, development and evaluation of their own or alternative explanations based on evidence and how these explanations are connected to scientific knowledge. These learning opportunities are expected to represent activities and strategies that involve students in intellectual work in a social environment that is safe for discussion and interaction. Criteria for defining the level of fidelity can include similarity between the

learning opportunities enacted and those intended in the curriculum and the quality of adaptations made to the learning opportunities. *Similarity* refers to the closeness between the intended activity with the enacted one in terms of the characterization of actions (dynamic) of teacher and students (e.g., during the small group interactions, do students interact as intended?) It should be expected that the measurement of similarity provides information about the quality of the learning opportunities to which students are exposed. *Quality of Adaptations* refers to adaptations consistent with learning goals and appropriate for students' needs.

The third component, *supporting student learning*, involves the teacher interventions directed to guide, examine, question, and shape students' thinking to improve their understanding and learning (Duschl, 2000, 2003; Kesidou & Roseman, 2002; Mortimer & Scott, 2000; Schneider, Krajcik, & Blumenfeld; 2005). I consider three main elements that have proved to be relevant in teaching and learning science in the context of science inquiry: guiding students learning, sharing students' ideas/understanding, and checking of student understanding. Table 1 presents some issues that can be considered in this component. Criteria for measuring level of fidelity in supporting students' learning can be the thoroughness or completeness of strategies and the appropriateness of strategies observed. It is important to mention that curricula hardly provide guidance for teachers on how to promote students' thinking or how to address commonly held student ideas (Kesidou & Roseman, 2002). However, the criteria provided can be aligned to the theoretical stand of the curriculum. *Thoroughness* refers to the completeness of all the appropriate strategies proposed in the curriculum and those enacted by the teacher. *Appropriateness* focuses on the suitability of the instructional support strategies used according to the characteristics of the content or activity at hand, and that match with the students' learning needs (Schneider, Krajcik, & Blumenfeld; 2005).

The measurement of the three components of instructional transactions help to collect information about quality of delivery mainly, but it also helps to make decisions on adherence and exposure. It is important to mention that one aspect of FOI, curriculum differentiation, is not linked to any of the components described in this section. The rationale behind this decision is the type of evaluation I had in mind when I was writing this paper. In the context of evaluation, effectiveness can be evaluated by *comparing* programs or by focusing on the issue of *degrees of implementation* – more or less of the intervention, in this case the curriculum (Berk & Rossi, 1990). Curriculum differentiation (à la Dane & Schneider, 1998) seems to be more appropriate when

comparing the effectiveness of diverse curricula. I focus on the second aspect, degree of implementation, and how it relates to the effectiveness of the curriculum.

Outcomes. The fourth curriculum dimension involves the knowledge and skills that the curriculum developers expect to achieve as a result of the instructional transactions. To measure outcomes, it is important to clearly specify what students are expected to know and do. Clearly specifying the type of knowledge expected based on the learning opportunities provided to the students helps to develop assessments that are more valid. One strategy to accomplish this goal is to conceptualize outcomes based on four types of knowledge (Li, 2001; Ruiz-Primo, 1997, 2002; Shavelson & Ruiz-Primo, 1999): *declarative knowledge (knowing that)* includes knowledge about scientific terms, definitions, facts, or statements (e.g., defining mass, volume or density, knowing that the water density is 1); *procedural knowledge (knowing how)* takes the form of if-then production rules or a sequence of steps (e.g., measuring mass using a balance, applying an algorithm to balance chemical equations, reading a data table, or designing an investigation); *schematic knowledge (knowing why)* entails the application of scientific principles or explanatory models (e.g., explaining why an object sinks in water and floats in alcohol); and *strategic knowledge (knowing when, where, and how)* to apply knowledge. This knowledge includes domain-specific strategies such as ways to represent a problem or strategies to deal with certain types of tasks. It also considers general monitoring performance or planning strategies (e.g., solving a new problem).

Based on the outcomes profile, student learning can be assessed using a suite of items that tap the different types of knowledge expected. Furthermore, these assessments can be of different proximities (close vs. distal) according to the characteristics of the learning opportunities (Ruiz-Primo, Shavelson, Hamilton & Klein, 2002).

Fidelity of Implementation: An Empirical Study

This section describes the general context in which the FOI study is embedded. Although data collection has been completed, the study is still in progress. Therefore, I provide a description of the strategies we used to collect the data and the instruments we are developing to capture degree of fidelity.

The Context

The FOI study is part of a larger project that focuses on the effects of formal embedded assessments on students' learning (Shavelson & Young, 2000). The study is

based on a small randomized experiment carried out over the 2003-2004 school year with six “experimental” and six “control” teachers. The study was conducted within the context of the “Foundational Approaches in Science Teaching” (FAST) middle-school science curriculum (Pottenger & Young, 1992).

FAST is a middle-school science education program developed by the University of Hawaii’s Curriculum Research & Development Group (CRDG) and is aligned with the National Science Education Standards (Rogg & Kahle, 1997). The program consists of three texts: FAST 1, *The Local Environment*; FAST 2, *Matter and Energy in the Biosphere*; and FAST 3, *Change over Time*.

The formative assessment study. Six matched pairs of FAST teachers were randomly assigned to experimental and control groups in a pre- and post-test design.¹⁴ The experimental teachers participated in a five-day training program focusing mainly on the implementation of formal embedded assessments that use assessment information to provide immediate feedback to students around the fundamental question that underlies the first 12 investigations of FAST: Why Do Things Sink and Float? We named the assessments Reflective Lessons rather than embedded assessments to make evident to teachers that their purpose was not to grade students. Reflective Lessons are a unique setting involving specific prompts designed for eliciting students’ conceptions, encouraging communication and argumentation, and helping students and teachers reflect about learning and instruction. The prompts vary according to where the Reflective Lessons are embedded within the unit. It is important to mention that the training program focused only on the implementation of the Reflective Lessons and not on the implementation of any of the Physical Investigations. This paper focuses on FOI of the FAST Investigations, but not on the Reflective Lessons.

The Formative Embedded Assessment project focused on the introductory Physical Science strand of *FAST 1, The Local Environment*. In this strand, students investigate concepts such as mass, volume, and density, as well as the relationship between density and buoyancy. In doing so, they work with different states of matter and use their knowledge to explain everyday phenomena. Within this strand, the study focused on the first twelve investigations (PS1 to PS12). Table 3 provides a quick summary of the characteristics of the twelve investigations (or units) and Figure 4 provides a schematic representation of where the Reflective Lessons (RL) are embedded within PS 1-12 (SEAL, 2003).

¹⁴ Ethnicity, free lunch, and student proficiency level were used to match the pairs as best as possible.

Table 3. Summary of the Characteristics of the Twelve Investigations

PS	Title of Investigation	Major Activities	Major Learning Goals
1	Liquids and Vials	Observing a buoyancy anomaly	Making scientific observations; testing predictions
2	Sinking a Straw	Adding mass to a straw and measuring its depth of sinking	Predicting the number of BB's required to sink a straw to a chosen depth
3	Graphing the Sinking-Straw Data	Creating a graph of mass versus depth of sinking	Representing data in graphs
4	Mass and the Sinking Straw	Sinking straws to depth based upon total mass	Increasing mass means a straw will sink more
5	Sinking Cartons	Sinking cartons of different sizes with equal mass	Predicting the depth to which a carton will sink
6	Volume and the Sinking Cartons	Calculating the volume of cartons	Calculating the displaced volume of a carton
7	Floating and Sinking Objects	Calculating the mass and volume of objects	Predicting the displaced water of floating and sinking objects
8	Introduction to the Cartesian Diver	Experimenting with a Cartesian diver	Discovering how a Cartesian diver works
9	Density and the Cartesian Diver	Finding the density of a Cartesian diver	Finding the density of a diver at different depths
10	Density of Objects	Calculating the density of objects	Finding the density of floating and sinking objects
11	Density of Liquids	Finding the density of liquids other than water	Finding the density of liquids
12	Buoyancy of Liquids	Finding the relationship between buoyancy and	Understanding relative density

First 12 Investigations of the FAST1 Physical Science Unit by Section																	
Section A Introduction to Buoyancy (Mass)					Section B Volume					Section C Density & Buoyancy							
1	2	3	4	RL	5	6	RL	7	RL	8	9	10	RL	11	RL	12	

Figure 4. Reflective Lessons Embedded in Across the Twelve Investigations (RL = Reflective Lesson)

Two questions guided the implementation study (Ruiz-Primo, 2003): (a) To what extent does the implementation of FAST and formative assessment reflect the objectives of content and pedagogy expressed in the FAST materials and the SEAL training program? (b) Does the quality of the implementation have an effect on student performance? In what follows, I focus only on the FOI of FAST and not on the implementation of the reflective lessons, or the relation to outcomes.

The Fidelity Study

Participants

Twelve teachers and the students participated in the Formative Assessment Study. Teachers were from 12 schools in 11 school districts from six states. A summary of the characteristics of the teachers is provided in Appendix A. Six of the teachers were part of the experimental groups (i.e., implementation of the Twelve FAST Physical Science Investigations and the Reflective Lessons) and the other six were in the control groups (i.e., implementation of the Twelve FAST Physical Science Investigations only).

Data Collection

Table 5 provides information on the data and the time in which it was collected. The instruments will be described in a later section. All teachers responded to a questionnaire and a set of vignettes before and after the implementation study. They were asked to fill in a web-based teacher log and to videotape themselves in every session they taught FAST; video cameras and videotapes were provided to each of them. Teachers were trained in the use of the video camera and the use of teacher logs prior to the beginning of the study. Teachers were asked to send weekly videotapes and any classroom artifacts used during that week (stamped envelopes were provided). Each classroom was visited once during the course of the implementation over a two- or three-day period. Although most visits were conducted during Investigation 7 (PS7), some classrooms were visited at a later time due to external factors (e.g., snow on the

East Coast). The purpose of the visit was twofold: to have a sense of the context in which teachers were instructing and to collect information on small groups. At the end of the school year, all the teachers provided their students' science notebooks.

Table 5. Type of Data Linked to Curriculum Implementation

Data Collected	Implementation Period		
	Before	During	After
Teacher Questionnaire	✓		✓
Vignettes	✓		✓
Teachers' Videotapes		✓	
Teacher Logs		✓	
Small Group Videotapes		✓	
Students' Notebooks			✓

Identifying FAST Components

As mentioned, the first step in measuring FOI should be the identification of the critical and related components of the curriculum at hand. FAST can be generally described as being based on a constructivist philosophy of learning, in which students construct their own knowledge and understanding from their experiences incrementally. This knowledge is developed and clarified through interactions with others (Pottenger & Young, 1992). Investigations are carefully sequenced and connected to previous experiences both in and out of school. Students often work in small groups to share data, ideas, and experiences; conduct investigations; summarize; and draw conclusions. Class discussion follows each investigation to identify and clarify generalizations.

The process of the identification of FAST components has not ended.¹⁵ Still, there seems to be an agreement on two critical FAST components: the role of the teacher as a discussion facilitator and the role that small groups play in constructing students' knowledge. However, a third element has been mentioned as pivotal to FAST: the carefully designed and sequenced investigations. The characteristics of the investigations are such that students are expected to overcome misconceptions as they advance in the conduction of the investigations. Also, the nature of the investigations

¹⁵ Defining the critical components, expected variations, and configurations of FAST has been an iterative discussion between the researchers and the CRDG curriculum developers, Frank Pottenger and Donald Young. It continues to be an ongoing process. Miki Tomita, a former and currently summer-only FAST teacher is fully involved in this process. However, the components presented in this paper were identified only by the author of this paper.

changes as the students advance in the unit. They start as guided investigations and are becoming increasingly more open as the students advance in the curriculum.

As an exercise for addressing the *intended curriculum*, Appendix B presents some FAST components based on the analysis of FAST materials that are available to teachers and students: Teacher Materials (Teacher's Guide and Instructional Guide) and the Student Materials (Student Book and Student Record Book).¹⁶ That is, only *public documents* were used as sources of information. The decision on which ones are critical and which ones are related is an issue to be discussed and determined at a later time.

Measuring FOI

Except for the *Site Context* component for which an instrument was not directly developed, the FOI study is measuring at least one aspect of all the components across curriculum dimensions presented in Appendix B. As mentioned before, teacher as facilitator of discussions and small groups is the only component that can be considered critical at this time.

We are using a multi-method and multi-source strategy to track the enacted and achieved curricula and the curriculum dimensions. Following the FOI instrument dimensions previously presented (see p. 11), the instruments used involved different levels of judgment (some are more objective than others), levels of directedness (some capture the implementation more directly than others), levels of sensitivity (some instruments are more sensitive than others), and levels of alignment (some instruments are more aligned to FAST than others). We used three sources of information, the curriculum providers (i.e., teachers), independent observers (i.e., the researchers), and the participants (i.e., students).¹⁷ The diverse sources of information allow triangulating the information and providing a good source of validation across instruments.

Figure 5 presents a summary of the instruments by data source linked to the curriculum dimensions and the aspects of fidelity. The figure also provides information on general characteristics of the instruments based on the four dimensions (extent of judgment, directedness, sensitivity, and alignment with FAST) used to describe the instruments. Based on the four dimensions and using a rough overall judgment according to how the instruments were developed, I classified the instruments using the following categories: Low (L), Medium Low (ML), Medium (M), Medium High (MH),

¹⁶ Reference Booklets were not reviewed.

¹⁷ The FOI study is a composite of sub-studies, one per instrument developed. Different researchers are involved on each of these studies.

and High (H). The question mark indicates that the judgment will result from the analyses in progress. The explanation of the figure is presented according to the components measured.

Measuring teaching beliefs and values. In this sub-study, conducted in collaboration with Noah Feinstein (from SEAL), two instruments were developed, an 89 item-Likert-type scale self-efficacy questionnaire (Pajares, 1996) and a free-response vignette questionnaire (Kennedy, 1999; Ruiz-Primo & Li, 2002, 2003).

The framework used to develop the instruments focused on three well-defined constructs. The first one is *Epistemological beliefs* – beliefs about the nature of knowledge (Schommer-Aikins & Hutter, 2002). We hypothesized that teachers would seek out and value different evidence of student understanding based on their epistemological beliefs. The second construct is *Outcome expectancies* – beliefs that a particular course of action will have a particular outcome (Eccles, 2002). We hypothesized that outcome expectancies would influence the actions teachers took based on their perceptions of student understanding. The third construct is *Self-efficacy* – confidence in one’s ability to successfully perform an action (Bandura, 1997). We hypothesized that teachers would choose instructional strategies that they felt capable of carrying out successfully.

Data Source	Instrument	Instrument Profile				Critical Component Tapped	FOI Aspect			
		Extent of Judgment	Extent of Directedness	Extent of Sensibility	Alignment with FAST		Adherence	Exposure	Quality of Delivery	Participant Responsiveness
Teacher	▪ Beliefs Questionnaire	L	L	?	ML	▪ Teachers beliefs and values				
	▪ Vignettes	M	L	?	H					
	▪ Teacher Log	M	ML	?	MH	▪ Providing Learning Opportunities ▪ Supporting Student Learning	✓		✓	
Teachers' Videotapes	▪ Investigation Maps	L	H	H	H	▪ Critical Topics And Activities		✓		
	▪ Questioning Strategies Maps	M	H	H	H	▪ Providing Learning Opportunities ▪ Supporting Student Learning	✓	✓	✓	✓
	▪ Formative Assess. Coding System	M	H	H	M		✓	✓	✓	
Small Group Videotapes	▪ Argumentation Coding System	M	H	H	H	▪ Supporting Student Learning	✓			
	▪ Investigation Group Map	M	H	H	H	▪ Developing & Using Scientific Knowledge	✓			✓
Classroom Artifacts	▪ Quality of Artifacts Coding System	M	L	M	MH		✓			
Students' Notebook	▪ Notebook Scoring System	M	MH	M	H		✓	✓	✓	✓
Student Performance	▪ Multiple-Choice Test					▪ Outcomes				
	▪ Predict-Observe-Explain									
	▪ Performance Assessment									
	▪ Short-Answer									

Figure 5. Instruments by Data Source, Critical Components, and Aspect of FOI

Of the 89 items on the questionnaire, 31 asked general information about teachers (i.e., background and professional development) and instructional practices (e.g., how often have you employed the following teaching strategies?). Five of these questions involve selection, 10 are short-answer, and 16 are Likert-type questions. The other 58 items consisted of four-point Likert-type scale (strongly agree to strongly disagree). The

items focused on three areas: *science inquiry* (e.g., I am effective in modeling to my students how to gather, record, analyze and interpret data), *formative assessment* (e.g., During class discussions, I can quickly identify important misconceptions students have about the material), and *general instructional strategies* (e.g., If I have enough time, I can find a good way of helping my students connect a new lesson to material they've already learned).

Four vignettes were developed focusing on teacher assessment practices. Vignettes are viewed as second-level approximation indicators to study factors that influence student learning, as opposed to first-level indicators, such as classroom observations (Kennedy, 1999). They provide descriptions that are as closely situated as possible to teachers' own practices (Ruiz-Primo & Li, 2002, 2003). They aim to tap teachers' actual practices more than generalities or vagaries. Because they are supposed to be closely related to classroom practice, they probe teachers' knowledge-in-action as well as their practical wisdom rather than content knowledge (Kennedy, 1999). Appendix C provides an example of a vignette.

We are currently analyzing the technical qualities of both instruments. Teachers' videotapes will allow us to investigate the relationship between the belief constructs and teachers' adoption and implementation of curriculum materials. A pre-post-test design will allow us to test the stability of individual teacher's beliefs.

Qualitative analysis of the vignette data, currently in progress, focuses on teachers' interpretations of the information presented (epistemological beliefs) and their recommendations (outcome expectancies). Preliminary results indicate considerable stability over the study period: all of the 12 teachers offered similar interpretations and recommendations at pre- and post-test. This finding agrees with existing evidence that teachers' beliefs stabilize after the first years of teaching (Richardson, 1996). Trends within teachers' responses to the vignettes suggest discrete epistemological orientations that could have predictive value if assessed in a more targeted, sophisticated manner. Further analysis will attempt to link outcome expectancies to distinct patterns in curriculum enactment, identified using the videotaped observations.

Measuring topics, activities, and sequence. The strategy for approaching the measurement of some of the FAST components involved a fundamental activity that I named, *mapping* the FAST investigations. With Erin Furtak and Miki Tomita (from SEAL), the content and activities were inventoried on *investigation maps*. This inventory

reflects the organization and sequence of each of the FAST investigations.¹⁸ Figure 6 provides a sample of a map. The letter in parentheses indicates who (teacher, students, small group) the main actor is involved in the content described in the map. The maps allow us to identify “other” activities implemented but not required by the curriculum. The maps followed the sequence and organization of the investigations.

Investigation maps are filled in by independent observers (researchers). The data source used for this purpose is the teachers’ videotapes. For exposure purposes, the investigation map leads to an index of the percentage of elements enacted. This index leads to a decision about the level of completeness: complete – all the important elements of the intended curriculum were enacted; sufficient – all the appropriate elements are addressed but some minor are missed; incomplete – some important elements are missing; and insufficient –several main ideas are missing.

Members participating T, S, T-S, G, C	Instructional Sequence	PS1. Liquids and Vials							
		<p><i>PS1 Objectives:</i> Sensitize students about science is concerned with actual phenomena Provide an introduction to the nature of science, search of explanations Identify buoyancy phenomena as the focus of investigation</p>							
		<i>Introducing FAST</i>							
		Discuss characteristics of materials and exercises (T)							
		Discuss student notebooks (T)							
		Discuss laboratory safety and procedures (T)							
		<i>Observing Liquids-and-Vials</i>							
		Show Liquids to students (T)							
		Invite student observations of Liquids (T)							
		Record observations on Liquids (T), (S)							
		Show Vials to students (T)							
		Discuss/Define <i>vials</i> (T)							
		.							
		.							
		.							
		<i>Other Activities</i>							

Figure 6. Basic skeleton of an investigation map. The map presents a portion of Investigation 1 (PS1).

¹⁸ The level of specificity of the maps is still an issue of discussion with the curriculum developers. They have reviewed and commented on the PS1 map.

The investigation maps are linked and adapted according to the curriculum component being measured. Therefore, the maps are thought of as the intersection between *content* (concept, activities and sequence) and diverse aspects according to the key components of the curriculum, say the questioning strategies. This characteristic of the map will become clearer on the next section.

Developing scientific knowledge, providing learning opportunities, and supporting student learning. Instructional transactions seem to be the core of FOI studies of science curricula. Four sub-studies are being conducted to capture instructional transactions, but only two of them focus on the two FAST critical components discussed previously, teacher as facilitator of discussions and small groups work. In what follows, I describe how these two critical components are being measured.

FAST describes the role of the *teacher as facilitator* who guides the discussion towards a desired direction. The FAST Instructional Guide provides “Guidelines for Leading Discussions” and recommends four types of questions that teachers can use to guide the discussions.¹⁹ Teachers’ questioning practices are being coded where discussions are expected to happen. More specifically, we have focused on the discussion at the end of the investigations. This sub-study, conducted with Miki Tomita, focuses on tracking teacher questioning practices using the *Questioning Strategies Map* (Figure 7). The map captures: (a) the target content – Summary Questions and Challenge Questions; (b) the degree of teacher facilitation to engage students – from No Facilitation to High Facilitation; (c) members participating – teacher, individual student, small group, or class; and (d) the questioning strategies grouped in patterns.

Teachers’ videotapes are used as the source for filling-in the questioning strategy maps. It helps to capture the activities implemented, sequence, and questions used. Other pieces of information can also be captured (e.g., goal for the class session). This information helps to make an overall judgment of the teacher as a facilitator on a particular day. Based on the information collected from the level of facilitation, participating members, questioning strategies, and enactment of all the elements for discussion, different level of completeness and similarity can be assigned to capture adherence, exposure, and quality of delivery. Completeness is captured in a four level-scale: From complete (all summary and challenge questions were enacted) to

¹⁹ Questioning strategies suggested are: *Prompt* - Prompts student for response, simple questions; *Clarify/Elaborate* - Elicits meaning of unfamiliar terms, rephrase disconnected or fuzzy phrases; pushes for elaboration, probing for additions, analogies, alternatives, explanations; *Lift* - Moves conversation from narrow to broad, from specific to general, from concrete to general; *Summarize* - Refocuses meaning of statements or activities.

insufficient (only one summary question was enacted). Similarity varies from highly similar (teacher guides the discussion in the desired direction and uses the appropriate questioning strategies) to non-similar (teacher does not discuss or teacher does not lead the discussion in the desired direction and questioning strategies are not appropriate).

It is important to note that the map helps to guide the decision about the quality of the facilitation. Complete agreement on the map is not expected and, therefore, not analyzed. We looked for consistency between coders at the criterion level. Tomita (2005) found that the inter-rater agreement on 30% of the observations piloted (20) was 0.90. She found that the map could differentiate questioning strategies and levels of completeness across teachers that were more or less similar to what was expected by FAST.

These maps are designed to show the content of the curriculum materials and their sequence in the classroom, who participates during the instructional activities, and the **teacher as facilitator**.

Teacher: _____ Observer: _____ Date: _____

Tape#: _____ Start point: _____ End point: _____
 Tape#: _____ Start point: _____ End point: _____
 Tape#: _____ Start point: _____ End point: _____

Menters participating: T, T-S, Q, T-G, T-C, S (Or HW) - to be completed as an assignment.		Level of Directedness	Instructional Sequence	PS1: Liquids and Mals				Characteristics of Questions:		
				Dichotomous Questions:	FAST Objectives - students are driven towards being able to	PSF Objectives:	CC - Some indicators:	Vertical Moments: Facilitations	Call only S's volunteers	
				Use symbolic tools of science	Y	N	Use symbolic tools of science	Y	N	
				Understand the nature and philosophy of science	Y	N	Understand the nature and philosophy of science	Y	N	
				Make connections among investigations and concepts	Y	N	Make connections among investigations and concepts	Y	N	
				Apply scientific knowledge	Y	N	Apply scientific knowledge	Y	N	
				Serifize students about science is concerned with actual phenomena	Y	N	Serifize students about science is concerned with actual phenomena	Y	N	
				Provide an introduction to the nature of science, search of explanations	Y	N	Provide an introduction to the nature of science, search of explanations	Y	N	
				Identify buoyancy phenomena as the focus of investigation	Y	N	Identify buoyancy phenomena as the focus of investigation	Y	N	
				Explicit Class Objectives?	Y	N	Technological applications?	Y	N	
				Connections were made?	Y	N	Practical applications discussion?	Y	N	
				Monitoring around groups?	Y	N	Raise cognitive conflict?	Y	N	
				(Primary) Mode of recording student responses:						
				Overhead transparency		Chart paper		Notebook		
				Whiteboard		Computer		None		
				Level of Directedness: NF (Non-Facilitator) to F (Facilitator)						
				4 Encourage discussion/participation among those Ss who volunteer AND also those who do not						
				3 Encourage discussion AND is handled successfully BUT call only on Ss who volunteer						
				2 Encourage participation BUT discussion not handled successfully (no justification, no interest in student opinions) OR discussion not successful						
				1 Does not encourage discussion/verbalization						
				Leaving is appropriate OR discussion is not necessary						
				DISCUSSION						
				What are the phenomena in the Liquids and Mals system that you can explain? (Sum Q 1a)						
				What is your hypothesis to explain the phenomena? (Sum Q 1a)						
				What are the phenomena in the Liquids and Mals system that you cannot explain? (Sum Q 1b)						
				What were the results from testing your hypotheses? (Sum Q 2a)						
				Did your results support your hypothesis? (Sum Q 2b)						
				OR How might you have tested one of your hypotheses? (Sum Q 3, replaces Sum Q 2a and 2b)						
				How would you sink a can or a jar in a bucket of water so that only half of it is submerged? Explain what you did and how you did it. (Challenge Q)						
				Additional Activities						

Figure 7. An example of the Questioning Strategies Map for Investigation 1 (PS1).

The second FAST critical component is small group work. FAST investigations are intended for teams of 2 to 4 students, allowing “each student to engage in the investigation” and ensuring “an intellectual critical mass” (Young & Pottenger, 1992, p. 41). Such groups are assumed to exhibit various qualities, including open communication and shared responsibility (Young & Pottenger, 1992). The rationale is that group work is an important contributor to student understanding. This shared meaning construction is particularly crucial in the context of scientific inquiry. Group work can facilitate “convergent conceptual change” in which group members construct shared meanings of concepts and experiences (Roschelle, 1992). This shared meaning construction is particularly crucial in the context of scientific inquiry, reflecting a unique form of socially situated reasoning and knowledge building (Cobb & Yackel, 1996).

This sub-study is being conducted with Colin Schatz. We are focusing on the following issues: (a) members engaged in conducting the investigations (e.g., in some groups all members are engaged, but in some others most members are not actively engaged or some are disruptive); (b) leadership (e.g., negotiation on who will be conducting the investigation, negotiations in decision making); and (c) teacher interactions with the small group (e.g., informal analyses indicate that small group work was consistently characterized by minimal teacher intervention in the groups’ social processes or sometimes this interaction is disruptive). To capture this information we have used the *Small Group Investigation Map* (Figure 8). The map captures some aspects of the teacher actions while monitoring the small group, a general judgment about different aspects of the group activity, utterances among group members, and some characteristics of the students (i.e., gender and leadership).

Different data sources are used in this study: (a) video data gathered over the course of two or three days of observation and including an average of five student groups for each of the 12 teachers; (b) teacher interviews that focused on how students were assigned to groups; and (c) a Likert-type scale questionnaire that gathered teachers’ impressions of each group member’s overall performance, social status, responsibility, leadership, and receptivity to others’ ideas. Independent observers code the small group videotapes.

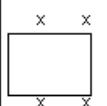
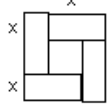
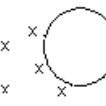
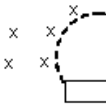
6. Volume and the Sinking Cartons											
<ul style="list-style-type: none"> How much volume is displaced by cartons of different sizes sinking to different depth in water? Can the volume or amount of displaced water be predicted if the volume of the part of a carton that is below the surface of the water is known? 											
Teacher		The group activity:									
Visits group?	Y N	Gives everyone the chance to join the discussion and/or contribute ideas?		Y N							
More than once?	#	Gives everyone opportunity to participate in decision-making?		Y N							
Talks to students		Gives everyone opportunity to communicate?		Y N							
Give questions that generate student engagement?	Y N S	Appears to be solving the problem collaboratively?		Y N S							
Clarify objectives and/or tasks?	Y N S	During the group activity:									
Guides Ss reasoning: Challenges their conceptions?	Y N S	Were hypotheses/predictions provided?		Y N S							
Guides Ss reasoning: Attends to their thinking?	Y N S	Were reasons or justifications provided?		Y N S							
Communicates expectations and standards for their work?	Y N S	Were explanations built?		Y N S							
Makes evaluative statements about what students are doing?	Y N S	General Judgments:									
Focuses on procedures?	Y N S	Time spent (percentage) in figuring out how to complete the task									
Request information only?	Y N S	All steps followed?		Y N							
Identifies group needs?	Y N S	Predictions before testing?		Y N S							
Offers help when needed?	Y N S	Careful measurement behavior?		Y N S							
Group Size		Group									
Open-Ended Discussion – 10 to 12?	Y N	Knows what to do since the beginning?		Y N							
Problem Solving – 4 to 6?	Y N										
Laboratory Problems – 2 to 4?	Y N										
Short Discussion – 2?	Y N										
Seating Configuration											
Sequence	A	B	C	D	Other	Gender					
							Leader				
							S1	S2	S3	S4	S5
Determining the Submerged Volume of Cartons											
<ul style="list-style-type: none"> - Measuring the base area of the carton (length x width) - Recording the base area (cm²) (table provided) - Determining the depth of sinking (rubber band around carton) - Recording the depth of sinking (table provided) - Calculating the submerged volume of the carton (area of base x depth of sinking) - Recording the submerged volume (table provided) 											
Sinking the Cartons											
<ul style="list-style-type: none"> - Adding sand to the cartons until it sinks at the depth marked - Measuring and recording the water displaced (Table provided) 											

Figure 8. Small group map for Investigation 6 (PS6).

In order to capture the informal formative assessment (IFA) practices, we (Ruiz-Primo & Furtak, 2004) developed the Informal Formative Assessment Coding System. The instrument captures aspects related to Supporting Students' Learning: conveying unit/lesson purpose, connecting ideas, modeling, providing feedback, sharing student ideas, and especially, checking student understanding. This coding system focuses on informal assessment practices in the context of science inquiry; therefore, the system is not aligned to FAST and does not use any type of investigation map. Information

collected with this instrument is similar to information yielded from the questioning strategies map. This can be considered as evidence of the convergent validity of the questioning strategies maps.

The sub-study on the use of teacher logs is being conducted with Kun Yuan (from SEAL) and Erin Furtak. The log is a questionnaire that asks teachers simple questions on diverse characteristics of their instruction on any given day. It is composed of 13 questions focusing on different aspects of instruction: investigation taught on a particular day (1 item), goals of today's class (3 items), emphasis given to diverse aspects of student performance (e.g. memorization, solving problems, perform procedures; 1 item), materials used (e.g., FAST or other materials; 1 item), organization of students in today's class (e.g., whole group, small group, individual, 3 items), instructional strategies (e.g., facilitation, modeling, lecturing) during diverse instructional episodes (e.g., introducing a new topic, conducting an investigation; 1 item), students activities during today's class (e.g., engagement in discussions, complete worksheets; 1 item), overall student engagement (1 item), and general comments about today's class (1 item). The teacher log was posted on a web with the help of the University of Hawai'i. Teachers could fill in the log every day using a password provided to each of them. Researchers had access to the logs everyday with information about when the logs were submitted. All teachers from the control and the experimental group participated in a training session about the use of teacher logs in which each of the items was discussed, as well as the definitions of terms used in each question. Finally, 11 interview protocol questions were developed to tap teachers' conceptions about items related to informal formative assessment practices in class (e.g., In your mind, what is a typical scenario for facilitating class discussion?)

Notebooks. In collaboration with Min Li (University of Washington) and Marsha Ing (UCLA) we are analyzing information from students' notebooks. The use of notebooks is encouraged by FAST and proposed as a textbook constructed by the student. We have scored notebooks across the 12 teachers. We sampled notebooks over four investigations. We developed a scoring system that focuses on the types of entries promoted by FAST: problem, vocabulary, background information, method, reporting results, producing explanations and drawing conclusions. Special attention is paid in the scoring system to the quality of the explanations provided by the students. We are exploring a computer-based scoring system to facilitate the process and to explore factors that may affect consistency between raters.

Finally, another sub-study conducted with Jonathan Dolle (from SEAL) focuses on the analysis of classroom artifacts collected during the implementation of FAST. The coding system focuses on ten dimensions: type of artifact, individual/collaborative artifact, source of artifact, alignment with FAST, type of response elicited from the student, type of cognitive activity elicited, level of guidance in the procedure, and identification of erroneous concepts or procedures in the design of the task. We are currently coding the artifacts across teachers and investigations.

Measuring Outcomes

As part of the Formative Students Study, students were administered a multiple-choice achievement test and a motivation questionnaire as pre-test and post-test assessments. In the post-test we also administered three more assessments of different types: a performance assessment, a predict-observe-explain assessment, and one open-ended question.

The 43-item multiple-choice test included questions on density, mass, volume, and relative density. Eight items focused on definitions (e.g., What is the mass of 1 cm^3 of pure water?) Of the remaining questions, almost half focused (18 items) on the application of students' factual and conceptual knowledge in "doing" science.

For example: Julie put a $4 \text{ cm} \times 4 \text{ cm} \times 4 \text{ cm}$ block into an overflow can. She finds that the volume of water displaced is 32 cm^3 . What is the block's submerged volume?

The rest of the questions (17 items) focus on explaining and reasoning using conceptual and procedural knowledge.

For example, Willy has three cartons. Their bottoms have different surface areas: Carton A = $6 \text{ cm} \times 6 \text{ cm}$, Carton B = $7 \text{ cm} \times 7 \text{ cm}$, and Carton C = $8 \text{ cm} \times 8 \text{ cm}$. He puts the SAME mass of BBs into each carton. Which carton will displace the greatest volume of water?

The internal consistency of the test is 0.86 (Yin, 2005). Content, instructional, and construct validity of the assessment has been established by aligning each item to the

content of the FAST investigations, comparing pre-test and post-test, and conducting a series of factor analyses (Yin, 2005).

The motivation questionnaire includes the following dimensions (Yin, 2005): goal orientation (alpha coefficient = 0.78), perceived context (0.78), epistemic beliefs (0.44), self-efficacy (0.81), science interest and values (0.89), self-reflection (0.76), ego approach (0.74), and peer review (0.69).²⁰

The performance assessment asks the student to conduct an investigation to find out the density of a mystery liquid. Students are provided with three blocks of different density, a bottle of water, a bottle of the mystery liquid, overflow can, and other materials they can use to carry out the investigation. The internal consistency across the different parts of the assessment is 0.81 and the averaged interrater reliability across six raters in two groups is 0.90 (Yin, 2005).

The predict-observe-explain assessment focuses on one event and engages students in three tasks. First, students *predict* the outcome of some event related to sinking and floating (e.g., Predict what will happen when the small piece of a soap is placed in water) and justify their prediction. Then students *observe* the teacher carry out the activity and describe the event that they see. And finally, students reconcile and *explain* any conflict between prediction and observation. The interrater agreement for this assessment is 92 percent (Yin, 2005).

The open-ended question involves one question that asks students for an explanation. For example: Explain why do things sink and float? Write or draw as much information as you need to explain your answer. Use evidence and examples to support your explanation.

The interrater agreement for this assessment averaged across three groups of scores is 87 percent (Yin, 2005).

Preliminary Results

Some of the instruments are still being piloted with random samples on the source of information available. Some have proved to be reliable, but evidence about their validity is yet to be determined. However, for some instruments, we have partial

²⁰ Some dimensions (e.g., goal orientation) include more than one construct (e.g., task goal, ego approach, and ego avoidance). Reported alpha coefficients are averaged across constructs.

evidence on this matter. Different instruments are providing similar profiles across teachers. For example, questioning strategies maps, the IFA coding system, and the teachers' logs concur on the profiles of those teachers that have been coded with the three instruments. When linking assessment practices with students' performance, those teachers who have practices more aligned with FAST critical components have shown the highest student performance on the different achievement assessments. Preliminary results on the instruments described can be summarized as follows. These results should be cautiously considered, since final evidence on reliability and validity is not available yet.

Questioning practices. Performance of teachers as facilitators of discussion clearly varies (Ruiz-Primo & Furtak, 2004; Tomita, 2005). Most of the teachers' questioning practices focus on prompting students for simple response, while a few teachers implement elaboration and lifting questions. We have found that the instructional transactions of teachers who are identified by the questioning strategies maps as having higher levels of similarity to the intended FAST critical components are those whose students have performed better on the diverse assessments used to measure student achievement.

Small group. Preliminary analysis indicates that small group work was consistently characterized by minimal teacher intervention in the groups' social processes. Patterns of individual engagement in group work were consistent with teachers' perceptions of students' social tendencies, and suggest that groups were generally dominated by higher-achieving students. Students' interactions focus mainly on procedures. Very few utterances have been coded as argumentative. Further analysis will examine the roles of gender, social status, and group composition as influences that shape the overall effectiveness of small groups and intra-group equity of engagement in the study.

Notebooks. Evidence about exposure to the investigations was found across all the teachers. However, differences in the quality of notebooks across teachers were immense. Despite the guide provided in FAST, the pieces of information found on notebooks were very different, from single pages to well-written and complete report of investigations. We paid special attention to the quality of students' explanations. Few classrooms showed the quality of explanation expected by FAST (e.g., evidence provided). We found that feedback provided by teachers varied across investigations and aspect evaluated.

Teacher logs. Preliminary analysis with four teacher's logs across two investigations (PS1 and PS4) showed that agreement between teachers and researchers'

logs for a particular lesson is higher amongst those teachers whose instructional practices are more aligned with FAST than with those whose practices are more distal (Ruiz-Primo, Yuan, Furtak, & Shavelson, 2005). Furthermore, students whose teachers' instructional practices were closer to the practices proposed by FAST showed a higher level of performance. However, higher level of performance was not consistent across all the assessments (Ruiz-Primo et al, 2005).

Once all instruments have been technically evaluated and all teachers have been coded, we will perform regression analysis to determine which components explain the variance of students' performance.

Final Comments

This paper proposes an approach for studying FOI of inquiry-based science curricula. The intention was to provide guidance about science curriculum components that should be considered for fidelity. However, the challenges for measuring fidelity in a reliable, valid, and practical manner still exist. The strategies proposed are currently being tested. However, preliminary results show that they are providing useful information of practical value. Information from the diverse instruments that we have developed has given us a portrait of the diverse ways in which FAST teachers are implementing the curriculum and how these forms of curriculum enactment affect student learning. Furthermore, the information that we are collecting with the instruments is helping us to better understand some of the factors that could affect the results of the Formative Assessment Study. For example, we did not observe significant differences between the experimental and control group. However, it cannot be concluded that embedded assessments are not effective since FAST fidelity of implementation is not warranted for those teachers in the experimental group (Yin, 2005).

If the instruments are found to be reliable and valid, some of the strategies proposed here can be improved through the accumulated experience of the use of these strategies in other projects. Furthermore, I envision the use of G theory for determining sampling issues over time and methods.

Lessons on FOI learned so far can be summarized as follows. First, determining the critical components of the curriculum is crucial for developing FOI instruments that as a group provide a thorough picture of the degree of implementation. Ambiguity about what contributes to the effectiveness of a program plays a key role. It not only decreases fidelity and transferability, but also makes it considerably more difficult to

develop strategies to measure fidelity. It leads to unclearness about the acceptability of adaptations and transformations before a curriculum cannot be considered the same one anymore. Some of the instruments developed in this study are not as aligned to FAST as they should be due to the lack of clarity on the critical components.

Second, program configurations are critical in the study of FOI. Variations in implementation are inevitable. Thus, in determining configurations of critical components, the degree of variation must be clearly specified. This task involves, among other things, specifying which variations will be considered as minor or major and which variations will be regarded as acceptable or unacceptable.

Third, planning FOI studies for a program during the time in which the program is actually being written appears to be the first step for a successful FOI study. Thinking about FOI may affect the project in unexpected ways. It may help to make critical program components more explicit. Defining the critical components is certainly not an easy task and it may affect other aspects of the curriculum (e.g., training, manuals, teacher guides). However, the likelihood for conducting FOI studies during the development of a curriculum is, unfortunately, low. Consequently, information on the critical and related components is hardly available. Therefore, if FOI information is needed for evaluation or research purposes, it is highly recommended that the FOI studies are designed concurrently with the evaluation or research projects. This will allow for multiple iterations around the identification of critical components, strategies, sources of information, instruments, and times for data collection.

Fourth, in our study we learned that direct observation seems to be unavoidable, but videotapes may not be a bad surrogate. Technical problems with videos, however, are also unavoidable; but overall, it is a good data source.

Finally, using multiple methods and multiple sources seems to be the most appropriate way to address all aspects of FOI.

References

- Alliance Access (1999). Data Driven Curriculum Reform. Vol.3, No. 3. http://ra.terc.edu/publications/Alliance_Access/Vol3-No3/data-driven.html.
- Bandura, A. (1997). *Self-efficacy. The exercise of control*. New York, NY: W. H. Freeman and Company.
- Bauman, L. J., Stein, R., E., K., & Ireys, H. T. (1991). Reinventing fidelity: The transfer of social technology among settings. *American Journal of Community Psychology*, 19(4), 619-639.
- Berk, R. A. & Rossi, P. H. (1990). *Thinking about program evaluation*. Newbury Park, CA: SAGE Publications.
- Berman. P. & McLaughlin, M. W. (1976). Implementation of Educational Innovation. *Educational Forum*, XL(3), 347-370.
- Blakely, C. Mayer, J. P., Gottschalk, R. G., Schmitt, N., & Davidson, W. S. (1987). The Fidelity – Adaptation debate: Implications for the implementation of public sector social programs. *American Journal of Community Psychology*, 15(3), 253-269.
- Boruch, R. F., & Gomez, H. (1977). Sensitivity, bias, and theory in impact evaluation. *Professional Psychology*, 8(4), 411-433.
- Cobb, P. & Yackel, E. (1996). Constructivist, emergent, and sociocultural perspectives in the context of developmental research. *Educational Psychologist*, 31, 175-190.
- Cook, T. D., Leviton, L. C., & Shadish, Jr. W. R. (1985). Program evaluation. In G. Lindzey & E. Aronson (Eds.) *Handbook of social psychology* (Vol. 1, pp. 699-777). New York: Random House.
- Dane, A. V. & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, 18(1), 23-45.
- Dobson, K. S., & Shaw, B., F. (1988). The use of treatment manuals in cognitive therapy: Experience and issues. *Journal of Counseling and Clinical Psychology*, 56(5), 673-680.
- Driver, R., Leach, J., Millar, R., & Scott, P. (1996). *Young's people images of science*. Buckingham, UK: Open University Press.

- Duschl, R. (2000). Making the nature of science explicit. In R. Millar, Leach, J., and J. Osborne (Eds.) *Improving science education. The contribution of research* (pp. 187-206). Buckingham, UK: Open University Press.
- Duschl, R. A. (2003). Assessment of inquiry. In J. M. Atkin & J. E. Coffey (Eds.) *Everyday assessment in the science classroom* (pp. 41-59). Washington DC. National Science Teachers Association Press.
- Dusenbury, L., Brannigan, R., Falco, M., & Hansen, W. B. (2003). A review of research on fidelity of implementation: Implications for drug abuse prevention in school settings. *Health Education Research. Theory and Practice*, 18(2), 237-256.
- Eccles, J. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53.
- Fullan, M. & Pomfret, A. (1977). Research on curriculum and instruction implementastion. *Review of Educational Research*, 47(1), 335-397.
- Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review*, 18(1), 37-50.
- Gresham, F. M., Gansle, K. A., Noell, G. H., Cohen, S., & Rosenblum, S. (1993). Treatment integrity of school-based behavioral intervention studies: 1980-1990. *School Psychology Review*, 22(2), 264-272.
- Harachi, T. W., Abbott, R. D., Catalano, R. F., Haggerty, K. P., & Fleming, C. B. (1999). Opening the black box: Process evaluation measures to assess implementation and theory building. *American Journal of Community Psychology*, 27(5), 711-731.
- Haynes, S. N. & O'Brien, W. H. (1990). Functional analysis in behavior therapy. *Clinical Psychology Review*, 10, 649-668.
- Hoolbrook, J. K., Gray, J., Fasse, B. B., Camp, P. J., & Kolodner., J. L. (2000). *Managing complexity in classroom curriculum implementation sites: Triangulating multi-level assessment and evaluation*. Paper presented at the AERA Annual meeting. New Orleans, New Orleans.
- Hord, S. M., Rutherford, W. L., Huling-Austin, L., & Hall, G. E. (1987). *Taking charge or change*. Alexandria, Virginia: Association for Supervision and Curriculum Development.
- Huntley, M. A. (2004, December). *Evaluation of traditional and reform mathematics curricula*. Presentation at the Applying Multiple Social Science Research Methods

- to Educational Problems: A National Forum. Washington, DC: The National Academies
- Kennedy, M. M. (1999). Approximations to indicators of student outcomes. *Educational and Evaluation and Policy Analysis*, 21(4), 345-363.
- Kennedy, M. M. (2004, April 7). Reform ideals and teachers' practical intentions. *Education Policy Analysis Archives*, 12(13). Retrieved [April 2, 2005] from <http://epaa.asu.edu/epaa/v12n13/>.
- Kesidou, S, & Roseman, J. E. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review. *Journal of Research in Science Teaching*, 39(6), 522-549.
- Li, M. (2001). *A framework for science achievement and its link to test items*. Unpublished dissertation at Stanford University.
- Lillehoj, C. J., Griffin, K. W., & Spoth, R. (2004). Program provider and observer ratings of school-based preventive intervention implementation: Agreement and relation to youth outcomes. *Health Education & Behavior*, 31(2), 242-257.
- Leithwood, K. A., & Montgomery, D. J. (1980). Evaluating program implementation. *Evaluation Review*, 4 (2), 193-214.
- Madaus, G. F., & Kellaghan, T. (1992). Curriculum evaluation and assessment. In P. W. Jackson (Ed.). *Handbook of research on curriculum* (pp. 119-154). New York, NY: Macmillan Publishing Company.
- Moncher, F. J., & Prinz, R. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review*, 11, 247-266.
- Mortimer, E., & Scott, P. (2000). Analysing discourse in the science classroom. In R. Millar, Leach, J., and J. Osborne (Eds.) *Improving science education. The contribution of research* (pp.126-142). Buckingham, UK: Open University Press.
- Mowbray, C., Holter, M. C., Teague, G. B., & Bybee, D. (2003). Fidelity criteria: Development, measurement, and validation. *American Journal of Evaluation*, 24(3),315-340.
- National Research Council. (1996). National science education standards. Washington DC: Author.
- National Research Council (2001). *Inquiry and the National Science Education Standards*. Washington, DC: National Academy Press.

- Oakes, J., Gamoran, A., & Page, R. N. (1996) Curriculum differentiation: Opportunities, outcomes, and meanings. In P. W. Jackson (Ed.). *Handbook of research on curriculum* (pp. 570-608). New York, NY: Macmillan Publishing Company.
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, 66 (4): 543-578.
- Pentz, M. A., Trebow, E. A., Hansen, W. B., MacKinnon, D. P., Dwywe, J. H., Johnson, C. A. (1990). Effects of program implementation on adolescent drug use behavior. The Midwestern Prevention Project (MPP). *Evaluation Review*, 14(3), 264-289.
- Pottenger, F. and Young, D. (1992). *The Local Environment: FAST 1 Foundational Approaches in Science Teaching*. University of Hawaii Manoa: Curriculum Research and Development Group.
- Richardson, V. (1986). The role of attitudes and beliefs in learning to teach. In J. Sikula (Ed.) *Handbook of research on teacher education*. (Second Ed.) (pp. 102-119) Washington, D.C.: American Educational Research Association.
- Roschelle, J. (1992). Learning by collaborating: convergent conceptual change. *The Journal of the Learning Sciences*, 2(3), 235-276.
- Rogg, S. & Kahle, J. B. (1997). *Middle level standards-based inventory*. Oxford, OH: Miami, University of Ohio.
- Ruiz-Primo, M. A. (1994). *Formative evaluation for teacher enhancement programs: An approach and case study*. Unpublished doctoral dissertation. University of California, Santa Barbara.
- Ruiz-Primo, M. A. (1997). Toward a framework of subject-matter achievement assessment. Unpublished Manuscript. Stanford, CA: Stanford University.
- Ruiz-Primo, M. A. (2002). On a seamless assessment system. Paper presented at the Seamless Science Education Symposium, AAAS Annual Meeting, Boston, MA.
- Ruiz-Primo, M. A., & Furtak, E. M. (2004). *Informal Assessment of Students' Understanding of Scientific Inquiry*. Paper presented at the American Educational Research Association Annual Conference, San Diego, CA
- Ruiz-Primo, M. A., Shavelson, R.J., & Baxter, G. P. (1995). *Evaluation of a prototype teacher enhancement program on science performance assessment*. In P. Kansanen (Ed.). *Discussions on Some Educational Issues.VI. Research Reports 145*. Department of Teacher Education, University of Helsinki.

- Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, 39(5), 369-393.
- Ruiz-Primo, M. A. & Li, M. (2002). *Vignettes As An Alternative Teacher Evaluation Instrument: An Exploratory Study*. Paper presented at AERA Annual Meeting. New Orleans, LA
- Ruiz-Primo, M. A. (2003). *On implementation and opportunity to learn*. Unpublished manuscript. Stanford Education Assessment Laboratory, Stanford University at Stanford, CA.
- Ruiz-Primo, M.A., & Li, M. (2003). *Assessing some aspects of teachers' instructional practices through vignettes: An exploratory study*. Paper presented at AERA Annual Meeting. Chicago, IL.
- Ruiz-Primo, M. A., Yuan, K., Furtak, E., & Shavelson, R. J. (2005, April). *On the validity of teacher logs as a source of information about informal classroom assessment practices*. Poster presented at the AERA annual meeting. Montreal, Canada.
- Scheirer, M. A. & Rezmovic, R. L. (1983). Measuring the degree of program implementation. *Evaluation Review*, 7(5), 599-633.
- Schmidt, W. H., Jorde, D., Cogan, L. S., Barrier, E., Gonzalo, I., Schimizu, K., Sawasa, T., Valverde, G. A., McKinght, C., Prawat, R. S., Wiley, D., Raizen, S. A., Britton, E. D., & Wolfe, R. G. (1996). *Characterizing pedagogical flow. An investigation of mathematics and science teaching in six countries*. Dordrecht. The Netherlands: Kluwer Academic Publishers.
- Schneider, R. M., Krajick, J., Blumenfeld, P. (2005). Enacting reform-based science materials: The range of teacher enactments in reform classrooms. *Journal of Research in Science Teaching*, 42(3), 283-312.
- Schommer-Aikins, M. & Hutter, R. (2002). Epistemological beliefs and thinking about everyday controversial issues. *Journal of Psychology*, 136(1): 5-20.
- Stanford Education Assessment Laboratory – SEAL (2003). *On The Integration Of Formative Assessment In Teaching And Learning with Implications for Teacher Education*. Paper presented at the at the EARLI 10Pth Biennial Conference, Padova, Italy.
- Shavelson, R. J., & Ruiz-Primo, M. A. (1999). On the assessment of science achievement. (English version) *Unterrichts wissenschaft*, 27(2), 102-127.

- Shavelson, R. J., & Young, D. (2000). *Embedding assessments in the FAST curriculum: On the beginning the romance among curriculum, teaching and assessment*. Proposal submitted at the Elementary, Secondary and Informal Education Division at the National Science Foundation.
- Snyder, J., Bolin, F., & Zumwalt, K. (1992). Curriculum implementation. In P. W. Jackson (Ed.). *Handbook of research on curriculum* (pp. 402-435). New York, NY: Macmillan Publishing Company.
- Tomita, M. (2005). Final paper for the doctoral course Education 274 at the University of California, Berkeley.
- Yin, Y. (2005). *The Influence of Formative Assessments on Student Motivation, Learning, and Conceptual Change*. Unpublished doctoral dissertation. Stanford University.
- Waltz, J., Addis, M. E., Koerner, K., & Jacobson, N. S. (1993). Testing the integrity of a psychotherapy protocol: Assessment of adherence and competence. *Journal of Consulting and Clinical Psychology, 61*(4), 620-630.
- Witt, J. C. & Elliot, S. N. (1985). Acceptability of classroom intervention strategies. In T. R., Kratochwill (Ed.). *Advances in School Psychology* (pp. 251-288). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

APPENDIX A

Information About the Participant Teachers and Their Classes*

Experimental						Control					
Teacher	Teaching Experience ^a	Degree	Student Grade Level	Number of Classes ^b	Average Class Size	Teacher	Teaching Experience	Degree	Student Grade Level	Number of Classes	Average Class Size
Alex	Total: 2 Science: 2 FAST: 2 FAST I: 2	BS MA	7	FAST: 5 Non-FAST: 1	29	Rache l	Total: 12 Science: 12 FAST: 12 FAST I: 48	BS MS	7	FAST: 3 Non-FAST: 0	28
Andy	Total: 5 Science: 5 FAST: 2 FAST I: 8	BA	7	FAST: 2 Non-FAST: 3	30	Lenny	Total: 6 Science: 3 FAST: 3 FAST I: 3	BS	7	FAST: 5 Non-FAST: 0	22
Becca	Total: 18 Science: 17 FAST: 12 FAST I: 12	BE	7	FAST: 5 Non-FAST: 0	21	Ellen	Total: 5 Science: 5 FAST: 5 FAST I: 5	BA Minor S	7	FAST: 6 Non-FAST: 0	27
Carol	Total: 23 Science: 10 FAST: 1 FAST I: 1	BA ME	6	FAST: 1 Non-FAST: 5	26	Sofia	Total: 28 Science: 15 FAST: 4 FAST I: 4	BS MA	6	FAST: 3 Non-FAST: 2	27
Danielle	Total: 3 Science: 1 FAST: 1 FAST I: 1	BA	6	FAST: 2 Non-FAST: 2	23	Ben	Total: 15 Science: 15 FAST: 11 FAST I: 9	BS MA	7	FAST: 5 Non-FAST: 1	26
Rob	Total: 14 Science: 14 FAST: 7 FAST I: 0	BS	7	FAST: 1 Non-FAST: 4	26	Seren a	Total: 22 Science: 6 FAST: 2 FAST I: 2	BS MA	6	FAST: 4 Non-FAST: 0	21

Note: * From Yin, 2005

(a) **Total**: the total years the teacher has been teaching. **Science**: the years the teacher has been teaching science. **FAST**: the years the teacher has been teaching FAST. **FAST I**: the times the teacher has been teaching **FAST I** (In some schools FAST I is taught more than once a year)?

(b) Number of classes indicates teachers' teaching load. Some teachers also taught non-FAST classes besides FAST classes.

APPENDIX B

Table 6 : FAST Components at a Global and Intermediate Level of Description (See Footnote 15)

Curriculum Dimensions	Critical Components
<p>I. Theoretical Stand</p> <ul style="list-style-type: none"> ▪ FAST is based on a constructive philosophy of learning (Young & Pottenger, 1992, p.2) ▪ Instructional Model: Inquiry 	<p>Site Context No information was found on this component on the public documents</p> <p>Teachers' beliefs and values</p> <ul style="list-style-type: none"> ▪ All learners construct their own knowledge and understanding from their experiences ▪ Knowledge is developed and clarified through interactions with others ▪ Learners must experience the joys and frustrations of advancing and testing hypotheses and submitting their work for peer critique ▪ Teachers as leaders and facilitators help the students work towards common goals or purposes
<p>II. Curriculum Materials</p>	<p>Content and Activities and Their Sequence</p> <ul style="list-style-type: none"> ▪ <i>Critical concepts and ideas</i> (Investigation 1 to 12) <ul style="list-style-type: none"> - Mass, volume, density, relative density, floating and sinking ▪ <i>Critical activities</i> <ul style="list-style-type: none"> - Investigations - Discussions - Summary Questions to help students organize, summarize, and make connections between investigations ▪ <i>Critical sequence</i> <ul style="list-style-type: none"> - Investigations are carefully sequenced and connected to previous experiences both in school and out of school to enable students to build their knowledge
<p>III. Instructional Transactions</p>	<p>Developing & Using Scientific Knowledge</p> <ul style="list-style-type: none"> ▪ Introduces new vocabulary words ▪ Uses key questions from the student book to guide class discussions ▪ Reviews concepts or skills learned in previous lessons relevant to current lesson <p>Providing Learning Opportunities</p> <ul style="list-style-type: none"> ▪ <i>Social and physical environment</i> <ul style="list-style-type: none"> - FAST teacher provides an attractive, inviting, open, accepting environment resulting in cooperation and mutual respect among all students (e.g., relevant bulletin board, information about on-going investigations, seating arrangement, adequate supplies) ▪ <i>Strategies</i> <ul style="list-style-type: none"> - Group Work <ul style="list-style-type: none"> * FAST relies on students working in research teams to generate the theoretical content of the program * Group interaction in planning and executing investigations, discussing and validating hypotheses, and summarizing and drawing conclusions is central * 70 to 80% of the students' time is spent at the laboratory (or field studies), 30% is devoted to data analysis, small group or class discussion, literature research, and report writing * Teacher is facilitator and students are researchers * Size of small group varies according to the task (open-ended discussion, problem-solving, laboratory problems, short discussion) - Class discussion that begins and ends each investigation is essential to identify and clarify generalizations hold in common. <ul style="list-style-type: none"> * Questioning strategies that tap recalling, processing, and applying. Types of questions (e.g., clarifying, extending, lifting) - Building Cognitive Conflict (e.g., comparing students explanations, feigning surprise at a faulty explanations or a correct one) - Peer teaching - Use of flow diagrams – pictorial representation of written directions or procedures ▪ <i>Instructional activities</i> <ul style="list-style-type: none"> - Conducting investigations - Notebook - Keep record and organized notes on observations, hypotheses, summary questions, and vocabulary

Continue

Table 6. Continues

Curriculum Dimensions	Critical Components
	<p>Supporting Student Learning</p> <ul style="list-style-type: none"> ▪ <i>Guiding students learning</i> - Teacher <ul style="list-style-type: none"> - Clearly define goals and objectives - Help students to organize, summarize, make connections between investigations, - Keep sight of short- and long-range objectives - Must relate parts of the program regularly <ul style="list-style-type: none"> * Between the investigations * Between subcomponents of evolving concepts * Use of concepts from one area of science in another area of science * Use of knowledge developed in science to guide social decisions - Is facilitator and coordinator of activities <ul style="list-style-type: none"> * Coordinate peer teaching - Focuses student’s attention on laboratory (and field) techniques, experimental design, hypothesis formation and provides adequate opportunity to complete assigned work - Engages students in self –evaluation - Attends to planning, executing and interpreting experiments and community validation of results ▪ <i>Sharing student’s ideas/understanding</i> - Teacher <ul style="list-style-type: none"> - Encourages mutual sharing of ideas and labor ▪ <i>Checking of student understanding</i> - Teacher <ul style="list-style-type: none"> - Is alert to the intellectual difficulties of students - Uses questioning strategies to further student, group, or class discussions and student understanding
<p>IV. Outcomes</p>	<p>Curriculum Level (Young & Pottenger, 1992, p. 3)</p> <p>Increase students’</p> <ul style="list-style-type: none"> ▪ “Capability to perform basic laboratory skills ▪ “Capability to use symbolic tools employed in science ▪ “Knowledge of concepts that are foundational to modern science ▪ “Capability to engage in scientific inquiry ▪ “Understanding of the sequential nature of the development of science ▪ “Understanding of the relationships among disciplines of science ▪ “Knowledge of the nature of science ▪ “Capability to apply scientific knowledge to other areas of endeavor ▪ “Understanding of the relationships among science, technology, and society ▪ “Capability to use scientific knowledge for making decisions ▪ “Internalization of the values, attitudes, and traits necessary for successful endeavor in science” <p>Unit Level (PS1-PS12)*</p> <p>Students will be able to</p> <ul style="list-style-type: none"> ▪ Create an intelligible and accurate table and graph from data ▪ Report that the greater the mass the greater the sinking, when volume is constant ▪ Indicate that mass is more universal than BBs, and define weight as the measure of stuff in matter ▪ Include volume as a predictor of depth of sinking--a carton sinks less if it is bigger given the same amount of ballast ▪ Indicate that displaced volume equals submerged volume ▪ Classify objects into Sinker and Floaters in water, given the mass and volume of objects ▪ Know that the mass of 1 cm³ of water is 1 g ▪ Predict displaced liquid volume based on the mass of a floating object ▪ Indicate that the mass of sinking objects is greater than the mass of the displaced water ▪ Indicate that mass and volume are related through density. Students know a density formula as the ratio between mass and volume (g/cm³) ▪ Use graph to calculate density ▪ Decide whether an object will sink or float based on its density ▪ Measure the density of objects by measuring their mass and the displaced volume of completely submerged objects ▪ Understand that different liquids have different densities ▪ Decide whether an object will sink, float, or subsurface float depends on its relative density relative to the medium’s density

* Learning outcomes defined by curriculum developers and researchers for the Formative Assessment Study.

APPENDIX C

Example of a Vignette

In what follows we've included four brief descriptions of FAST teaching scenarios, which we're calling "vignettes." Please read through each vignette and write a few sentences in response to the questions that follow.

We know that the questions are really broad, and that you could probably fill a book with your answers! Mostly, we're interested in "the bottom line" – whatever you feel is most important in each circumstance.

You will probably also feel like you don't have enough information to answer the questions. Do the best you can, based on the information that you have, and on *your own experience* as a FAST teacher. Remember that we're not looking for any particular type of answer – we want to know how *you* think about teaching FAST.

Vignette 3

You have just finished teaching "Mass and the Sinking Straw," in which students are asked to synthesize what they've learned about mass and sinking. About two-thirds of the class had no difficulty with this lesson: they were either able to make accurate predictions about how many BBs it would take to sink a straw, or explain why their inaccurate predictions were wrong. The rest of the class, including some of the students who normally take more time to learn this type of thing, had trouble making predictions, and seemed unable to explain why their predictions were wrong. Your schedule is tight, and you're worried about fitting everything in, but you also know that this material is important for the rest of the unit.

Questions Asked to the Teachers:

- What do you think are the most important arguments for moving on to new material in this situation, and why do you think so?
- What do you think are the most important arguments against moving on to new material in this situation, and why do you think so?
- Based on what you know, would you move on to new material or review? Given your decision, what might you do to foster understanding among the students who appear to need help?